EUREKAH.COM

MOLECULAR BIOLOGY INTELLIGENCE UNIT

Dmitry Nurminsky

Selective Sweep

NURMINSKY

MBIU

Selective Sweep

LANDES BIOSCIENCE

KA/PP

KLUWER ACADEMIC / PLENUM PUBLISHERS

LANDES BIOSCIENCE

VISIT…

MOLECULAR BIOLOGY INTELLIGENCE UNIT

Landes Bioscience, a bioscience publisher, is making a transition to the internet as Eurekah.com.

INTELLIGENCE UNITS

Biotechnology Intelligence Unit
Medical Intelligence Unit
Molecular Biology Intelligence Unit
Neuroscience Intelligence Unit
Tissue Engineering Intelligence Unit

The chapters in this book, as well as the chapters of all of the five Intelligence Unit series, are available at our website.

EUREKAH.COM

ISBN 0-306-48235-5

KLUWER ACADEMIC / PLENUM PUBLISHERS
233 Spring Street, New York, New York 10013-1578

LANDES BIOSCIENCE

PRINTED IN U.S.A.

EUREKAH.COM

NURMINSKY

MBIU

Selective Sweep

LANDES BIOSCIENCE

KA/PP

MOLECULAR BIOLOGY INTELLIGENCE UNIT

Selective Sweep

Dmitry Nurminsky, Ph.D.
Department of Anatomy and Cellular Biology
Tufts University School of Medicine
Boston, Massachusetts, U.S.A.

LANDES BIOSCIENCE / EUREKAH.COM
GEORGETOWN, TEXAS
U.S.A.

KLUWER ACADEMIC / PLENUM PUBLISHERS
NEW YORK, NEW YORK
U.S.A.

SELECTIVE SWEEP

Molecular Biology Intelligence Unit

Landes Bioscience / Eurekah.com
Kluwer Academic / Plenum Publishers

Copyright ©2005 Eurekah.com and Kluwer Academic / Plenum Publishers

All rights reserved.

No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the publisher.

Printed in the U.S.A.

Kluwer Academic / Plenum Publishers, 233 Spring Street, New York, New York, U.S.A. 10013
http://www.wkap.nl

Please address all inquiries to the Publishers:
Landes Bioscience / Eurekah.com, 810 South Church Street, Georgetown, Texas, U.S.A. 78626
Phone: 512/ 863 7762; FAX: 512/ 863 0081
http://www.eurekah.com
http://www.landesbioscience.com

ISBN: 0-306-48235-5

Selective Sweep edited by Dmitry Nurminsky, Landes / Kluwer dual imprint / Landes series: Molecular Biology Intelligence Unit

While the authors, editors and publisher believe that drug selection and dosage and the specifications and usage of equipment and devices, as set forth in this book, are in accord with current recommendations and practice at the time of publication, they make no warranty, expressed or implied, with respect to material described in this book. In view of the ongoing research, equipment development, changes in governmental regulations and the rapid accumulation of information relating to the biomedical sciences, the reader is urged to carefully review and evaluate the information provided herein.

Library of Congress Cataloging-in-Publication Data

Selective sweep / [edited by] Dmitry Nurminsky.
p. ; cm. -- (Molecular biology intelligence unit)
Includes bibliographical references and index.
ISBN 0-306-48235-5
1. Mutation (Biology) 2. Population genetics. 3. Natural selection. I. Nurminsky, Dmitry. II. Series: Molecular biology intelligence unit (Unnumbered)
[DNLM: 1. Selection (Genetics) 2. Evolution, Molecular. 3. Mutation. QH 455 S464 2005]
QH460.S454 2005
576.5'49--dc22

2004023424

CONTENTS

Preface ..... **vii**

1. **Inferring Evolutionary History through Inter- and Intraspecific DNA Sequence Comparison: The *Drosophila janus* and *ocnus* Genes** **1**
John Parsch, Colin D. Meiklejohn and Daniel L. Hartl
Detecting Selection by Inter- and Intraspecific DNA Sequence Comparison 1
Molecular Evolution of the *janus* and *ocnus* Genes in the *D. melanogaster* Species Subgroup 3
DNA Sequence Polymorphism in the *janus-ocnus* Region of *D. simulans* 5
Distinguishing between Demographics and Selection 8
Identifying Specific Targets of Positive Selection 10

2. **Rapid Evolution of Sex-Related Genes: Sexual Conflict or Sex-Specific Adaptations?** **13**
Alberto Civetta and Rama S. Singh
Are Males Conflicting with Females or Coadapting to Them? 14
Conflict Scenarios: Balancing Selection in Rapidly Evolving Immune System Genes 15
External Fertilization: How Are Marine Invertebrate Sperm Surface Proteins Different from Immune System Genes? 16
Internal Fertilization: Lessons from *Drosophila* 16
Arm-Race vs. Sex-Specific Coadaptation: Test of Hypotheses 17

3. **Selective Sweep in the Evolution of a New Sperm-Specific Gene in Drosophila** **22**
Rob J. Kulathinal, Stanley A. Sawyer, Carlos D. Bustamante, Dmitry Nurminsky, Rita Ponce, José M. Ranz and Daniel L. Hartl
The Origin of *Sdic* 24
The Molecular Structure of *Sdic* 25
Reduced Polymorphism in the Region of *Sdic* 26
The Issue of Background Selection 26
Further Evidence for a Selective Sweep 27
Bayesian Analysis of Polymorphism and Divergence in the *Sdic* Region 28
Rapid Evolution of Male-Specific Genes 29

4. **Detecting Selective Sweeps with Haplotype Tests: Hitchhiking and Haplotype Tests** **34**
Frantz Depaulis, Sylvain Mousset and Michel Veuille
Available Haplotype Tests 37
Alternative Hypotheses 40
Conditioning on S vs. θ 40
Intragenic Recombination 42
Sampling Strategy and Sliding Window 42
Power 43

5. **A Novel Test Statistic for the Identification of Local Selective Sweeps Based on Microsatellite Gene Diversity 55**
Christian Schlötterer and Daniel Dieringer
Material and Methods 56
Results 57
Discussion 63

6. **Detecting Hitchhiking from Patterns of DNA Polymorphism 65**
Justin C. Fay and Chung-I Wu
Reduction in Levels of Variation 65
Skew in the Frequency Spectrum 67
Linkage Disequilibrium 70
Population Subdivision and Changes in Population Size 72
Distinguishing Background Selection and Hitchhiking in Regions of Low Recombination 74

7. **Periodic Selection and Ecological Diversity in Bacteria 78**
Frederick M. Cohan
The Nature of Recombination in Bacteria 79
The Effect of Rare Recombination on Diversity within a Population 80
The Origins of Permanent Divergence 80
Effects of Periodic Selection beyond the Boundaries of the Ecotype 83
Periodic Selection and Discovery of Bacterial Ecotypes 86
Has Periodic Selection Occurred in Nature? 89

8. **Distribution and Abundance of Polymorphism in the Malaria Genome 94**
Stephen M. Rich

9. **Selective Sweeps in Structured Populations—Empirical Evidence and Theoretical Studies 104**
Thomas Wiehe, Karl Schmid and Wolfgang Stephan
Experimental Evidence 105
Theoretical Studies 110
Discussion 113

Index 119

EDITOR

Dmitry Nurminsky
Department of Anatomy and Cellular Biology
Tufts University School of Medicine
Boston, Massachusetts, U.S.A.
E-mail: dmitry.nurminsky@tufts.edu
Chapter 3

CONTRIBUTORS

Carlos D. Bustamante
Department of Biological Statistics and Computational Biology
Cornell University
Ithaca, New York, U.S.A.
Chapter 3

Alberto Civetta
Department of Biology
University of Winnipeg
Winnipeg, Manitoba, Canada
E-mail: a.civetta@uwinnipeg.ca
Chapter 2

Frederick M. Cohan
Department of Biology
Wesleyan University
Middletown, Connecticut, U.S.A.
E-mail: fcohan@wesleyan.edu
Chapter 7

Frantz Depaulis
Laboratoire d'Ecologie
EPHE-CNRS UMR
Université Pierre-et-Marie Curie
Paris, France
E-mail: fdepauli@snv.jussieu.fr
Chapter 4

Daniel Dieringer
Institut für Tierzucht und Genetik
Wien, Austria
Chapter 5

Justin C. Fay
Department of Genetics
Washington University School of Medicine
St. Louis, Missouri, U.S.A.
E-mail: jfay@genetics.wustl.edu
Chapter 6

Daniel L. Hartl
Department of Organismic and Evolutionary Biology
Harvard University
Cambridge, Massachusetts, U.S.A.
Chapters 1, 3

Rob J. Kulathinal
Department of Organismic and Evolutionary Biology
Harvard University
Cambridge, Massachusetts, U.S.A.
E-mail: rkulathinal@oeb.harvard.edu
Chapter 3

Colin D. Meiklejohn
Department of Organismic and Evolutionary Biology
Harvard University
Cambridge, Massachusetts, U.S.A.
Chapter 1

Sylvain Mousset
Laboratoire d'Ecologie
EPHE-CNRS UMR
Université Pierre-et-Marie Curie
Paris, France
Chapter 4

John Parsch
Department Biologie II
University of Munich (LMU)
Munich, Germany
E-mail:
parsch@zi.biologie.uni-muenchen.de
Chapter 1

Rita Ponce
Department of Organismic
and Evolutionary Biology
Harvard University
Cambridge, Massachusetts, U.S.A.
Chapter 3

José M. Ranz
Department of Organismic
and Evolutionary Biology
Harvard University
Cambridge, Massachusetts, U.S.A.
Chapter 3

Stephen M. Rich
Division of Infectious Disease
Tufts University School of Veterinary
Medicine
North Grafton, Massachusetts, U.S.A.
E-mail: stephen.rich@tufts.edu
Chapter 8

Stanley A. Sawyer
Department of Mathematics
Washington University
St. Louis, Missouri, U.S.A.
Chapter 3

Christian Schlötterer
Institut für Tierzucht und Genetik
Wien, Austria
E-mail:
christian.schloetterer@vu-wien.ac.at
Chapter 5

Karl Schmid
Max Planck Institut für Chemische
Ökologie and Jena Center
for Bioinformatics (JCB)
Jena, Germany
Chapter 9

Rama S. Singh
Department of Biology
McMaster University
Hamilton, Ontario, Canada
Chapter 2

Wolfgang Stephan
Department Biologie II
Sektion Evolutionsbiologie
Ludwig-Maximilians-Universität
München
München, Germany
Chapter 9

Michel Veuille
Laboratoire d'Ecologie
EPHE-CNRS UMR
Université Pierre-et-Marie Curie
Paris, France
Chapter 4

Thomas Wiehe
Institut für Molekularbiologie
und Biochemie
Freie Universität Berlin and Berlin
Center for Genome Based
Bioinformatics (BCB)
Berlin, Germany
E-mail: twiehe@zedat.fu-berlin.de
Chapter 9

Chung-I Wu
Department of Ecology and Evolution
University of Chicago
Chicago, Illinois, U.S.A.
Chapter 6

PREFACE

Positive selection is the driving force for the adaptation of organisms to an ever-changing environment, and it leads to adaptive evolution and in some cases to speciation. When selective pressure is applied to individuals based on their phenotype, it ultimately leads to the changes in the underlying genetic content of the population. The creatures that carry a more useful genotype would outcompete their peers, resulting in the fixation of beneficial allele(s) in the population with concomitant removal of inferior alleles. This process of selective sweep extends positive selection to the nucleotide level and therefore comprises the essence of Darwinian evolution. The genes that are subject to selection are usually found in the context of a chromosome. The adjacent genomic segments are physically linked to the selected genes and are therefore dragged to fixation along with the beneficial allele, or are discarded with the less fit alleles during the process called genetic hitchhiking. In some organisms, recombination can eventually separate the selected allele from adjacent loci; hence the strength of hitchhiking decreases with the distance from the selected locus. When recombination rates are very low, hitchhiking can drag to fixation extended regions of the genome or even entire chromosomes. In bacteria, the whole haploid genome represented by a single chromosome is driven to fixation during hitchhiking, which leads to rapid differentiation and speciation.

In the three decades since the description of the hitchhiking effect in the pioneering works of J. Maynard Smith and J. Haigh, and T. Ohta and M. Kimura, the issue has received variable attention. Discovery of low polymorphism in low recombination regions of the genome, consistent with the hitchhiking model, brought it into the spotlight – only until an alternative mechanism that explains the observed polymorphism pattern, the background selection, was proposed by B. Charlesworth. It became clear that unambiguous identification of a selective sweep event and associated hitchhiking is a formidable task. The complete selective sweep needed to induce a strong hitchhiking effect is associated with quite powerful positive selection – a rather rare case that is not readily found. In addition, the footprint of hitchhiking left by selective sweep on the pattern of polymorphism is initially not easily discernible from the pattern created by alternative mechanisms, such as background selection. Accumulation of new mutations after the sweep creates a distinctive signature of hitchhiking, but ironically the same mutation process quickly erodes the characteristic pattern. This creates a rather narrow time window for the detection of already rare strong hitchhiking event. Despite the described hardships, a number of selective sweeps have been documented, including the reports presented in this book. These reports play an important role in establishing the general framework of research on selective sweep and hitchhiking because they demonstrate that

theoretical predictions on polymorphism patterns derived from mathematical modeling are consistent with the real experimental data. However, analysis of these isolated examples largely leaves aside questions such as how often do the sweeps occur and what is the relative input of the sweeps and associated hitchhiking in shaping the pattern of polymorphism in the genome.

Recent advances in theory and technology have created a background for genome-wide surveys for selective sweep events. These advances include the development of new statistical tests tailored to detect incomplete, or partial, selective sweeps associated with weaker selection, and large-scale acquisition of DNA sequence data which provide ample source for the detection of polymorphism patterns. Whole-genome transcriptional analysis using gene microarrays is also instrumental in identification of male-specific genes, immunity genes, and other potentially rapidly evolving genes which are likely subject to significant positive selection and therefore represent probable targets for selective sweeps. Finally, detection of selective sweeps in structured populations has been significantly enhanced by theoretical analysis and the introduction of microsatellite polymorphic markers. While the complete picture is still emerging, it has become evident that selective sweeps and associated hitchhiking play the principal role in shaping the variability of the genomic sequence, and in selection-driven differentiation between populations.

The first three chapters, written by J. Parsch et al, R. Kulatinal et al, and A. Civetta and R. Singh, describe examples of selective sweep of rapidly evolving genes. Next three chapters, by F. Depaulis et al, C. Schlotterer, and J. Fay and C.-I. Wu, provide a comprehensive synopsis of the statistical methods for the detection of selective sweeps based on DNA polymorphism data, and also present original test statistics. The chapter written by F. Cohan provides an astonishing overview of the role of selective sweep in speciation in bacteria, and the chapter by S. Rich discusses the role of selection in shaping the variability patterns in the genome of the malaria plasmodium. Finally, T. Wiehe et al provides a very fine analysis of selective effects in structured populations.

My sincerest acknowledgements extend to the authors who contributed their time and effort to this book. It was a magnificent experience to work with these renowned scientists who were keen to share their great expertise with the readers. In addition, it was quite educational for me to edit their manuscripts, and I hope that the readers of this book will find it as useful and exciting as I did.

Dmitry Nurminsky

CHAPTER 1

Inferring Evolutionary History through Inter- and Intraspecific DNA Sequence Comparison:

The *Drosophila janus* and *ocnus* Genes

John Parsch, Colin D. Meiklejohn and Daniel L. Hartl

Abstract

Statistical analysis of aligned DNA sequences, both among and within species, has proven to be a valuable tool for inferring the evolutionary history of genetic loci. Of particular interest are cases where the observed data depart from the neutral expectation and suggest adaptive evolution due to positive natural selection. In this chapter, we use the *Drosophila janusA, janusB* and *ocnus* genes to demonstrate methods of evolutionary inference from both inter- and intraspecific DNA sequence data. Interspecific comparisons suggest that these three paralogous, testes-expressed genes have diverged in function following duplication and have evolved under different selective constraints. The three genes show the increased rate of between-species amino acid replacement common to genes with reproductive function, which may be the result of recurrent positive selection. Intraspecific comparison of *D. simulans* alleles provides evidence for more recent positive selection in this region of the genome. There are two divergent haplotype groups segregating in the worldwide population, one of which has risen to high frequency within the past 5000 years. The observed pattern of within-species variation may best be explained by a selective sweep that has not gone to completion.

Detecting Selection by Inter- and Intraspecific DNA Sequence Comparison

Recent advances in DNA sequencing technology have lead to an enormous increase in the amount of DNA sequence available for testing evolutionary hypotheses. This wealth of data includes collections of homologous gene sequences from different species as well as multiple sequences of particular genes sampled from different individuals within a single species. These data clearly indicate that there are abundant changes in DNA sequence between species as well as large amounts of DNA sequence polymorphism within species. Kimura's[1] neutral theory of molecular evolution explains this observation by assuming that the vast majority of nucleotide polymorphisms within species are the result of neutral mutations that have risen to detectable frequency due to random genetic drift in finite populations. Under Kimura's model, sequence differences between species reflect neutral polymorphisms that have drifted to fixation in one or the other species. Because evolutionary geneticists are primarily interested in the molecular basis of adaptive evolution, they focus largely on cases where the data depart from the neutral

Selective Sweep, edited by Dmitry Nurminsky. ©2005 Eurekah.com and Kluwer Academic/Plenum Publishers.

model. Such departures from neutrality may be detected through both interspecific and intraspecific DNA sequence comparison. Because divergence and polymorphism have a simple relationship under the neutral theory, the power to detect departures from neutrality is often increased by combining both inter- and intraspecific studies.

Interspecific DNA sequence comparisons are useful for determining the evolutionary history of particular genes and for identifying functionally important regions of genes and genomes. For example, interspecific comparisons can be used to estimate the order and timing of gene duplication events. They may also be used to identify functionally important protein motifs or gene regulatory sequences, which are expected to be highly conserved among species. One parameter that can be estimated from interspecifc DNA sequence data is the ratio of the nonsynonymous substituton rate (*Ka*) to the synonymous substitution rate (*Ks*). The *Ka*/*Ks* ratio (sometimes designated as d_N/d_S or ω) can reflect the selective constraints on a gene, particularly those acting to remove amino acid replacement substitutions. *Ka*/*Ks* = 1 is expected for genes evolving neutrally, where selection neither favors nor disfavors changes in the amino acid sequence. *Ka*/*Ks* < 1 is the commonly observed situation and suggests negative (purifying) selection acting to remove amino acid replacements. *Ka*/*Ks* > 1 indicates positive selection favoring the fixation of amino acid replacements. *Ka*/*Ks* > 1 is a strict criterion for the detection of positive selection and is rarely observed.[2] Notable exceptions include the antigenic proteins of some pathogens,[3,4] which are under strong selection to evade the host's immune response, and some male reproductive proteins that may be subject to sexual selection.[5-7] Recently, a number of maximum likelihood-based approaches for estimating *Ka*/*Ks* ratios for particular protein regions or amino acid positions have been introduced[4,8] that have increased statistical power to detect signatures of positive selection from interspecific data, particularly when a wide sampling of species with a known phylogenetic relationship is available.

Intraspecific DNA sequence comparisons can allow the detection of recent positive selection, that is, selection acting much more recently than the time of the last speciation event. For example, one can infer selection by a departure from the neutral expectation in the average number of nucleotide differences between two sampled alleles, also known as nucleotide heterozygosity. Positive directional selection is expected to cause a reduction in nucleotide heterozygosity in the genomic region linked to the selected site. This is because as the selected variant increases in frequency in the population and eventually goes to fixation, it will drag linked neutral variants to fixation along with it. Thus this phenomenon, known as genetic hitchhiking, or a selective sweep, leads to a decrease in the standing level of DNA polymorphism.[9] The extent of the chromosomal region affected by a selective sweep depends on the strength of selection and the local recombination rate.[10,11] The lower the recombination rate, or the stronger the selection, the larger the region of the genome that will be affected. The observation that chromosomal regions with little or no recombination show reduced levels of polymorphism when compared to regions of normal recombination in *Drosophila*[12-15] is consistent with genetic hitchhiking having acted in these areas. However, it has also been proposed that this observation can be explained by recurrent purifying selection removing linked neutral variants from the population. This mechanism, known as background selection,[16] is also expected to be stronger in regions of reduced recombination.

A number of statistical tests have been proposed to detect departures from neutrality using only intraspecific polymorphism data. For example, the test of Tajima[17] compares the observed frequencies of variants at polymorphic sites to the frequencies expected under the neutral theory. Other tests, commonly referred to as haplotype tests, compare the distribution of variants at segregating sites among chromosomes within a population sample to the neutral expectation.[18,19] Significant departures from neutrality detected by either Tajima's or the haplotype tests may be attributable to various forms of natural selection or to demographic factors reflecting the historical size and geographic distribution of the population. For example, an

excess of low frequency variants detected by Tajima's test can result from either a recent selective sweep or from recent population expansion. Typically, it is impossible to distinguish among these possibilities from analysis of a single locus, and data from additional loci are required before a conclusion can be reached.

The combination of inter- and intraspecific DNA sequence comparisons can be used for more powerful statistical methods of detecting departures from neutrality. In addition to DNA sequence polymorphism data from within a single species, the presence of at least one sequence from a closely related species can be of great value for two reasons. First, the DNA sequence from the related species can be used to classify the within-species DNA polymorphisms as derived or ancestral. In this case, the ancestral variant at the polymorphic site is assumed to be the one that matches the outgroup sequence. Examples of neutrality tests that consider the frequency of derived variants include those of Fu and Li,[20] Fay and Wu,[21] and Kim and Stephan.[22] The latter two tests are particularly relevant to genetic hitchhiking, because hitchhiking with a positively selected variant is expected to increase the frequency of derived variants in a population sample. The second benefit of having at least one sequence from a closely related species is that the divergence data can be used to eliminate variation in mutation rates or selective constraints as a cause for between-locus (or within-locus) differences in intraspecific polymorphism. For example, the low levels of DNA sequence polymorphism observed in regions of low recombination in *Drosophila* could be explained in theory by relatively low mutation rates in these regions of the genome. The observation that genes in regions of low recombination do not show a correlated reduction in interspecific divergence eliminates this possibility.[23] The expected correlation between divergence and polymorphism has lead to the development of several statistical tests of neutrality, including the HKA test[24] and the MK test.[25] The latter test compares ratios of polymorphism and divergence between synonymous and nonsynonymous sites from within a single protein-encoding gene. An excess of nonsynonymous changes between species can occur as a result of positive selection for amino acid replacements, although there may be other causes for this pattern as well.

In this chapter, we use the *Drosophila janus* and *ocnus* genes to illustrate the utility of inter- and intraspecific DNA sequence comparison for inferring evolutionary history. Several features of these genes make them interesting for evolutionary studies. First, they are a group of paralogous genes that have been created by several gene duplication events, with apparent specialization and functional divergence following duplication. Second, they are male-specific, testis-expressed genes that show the increased rate of molecular evolution characteristic of many *Drosophila* genes with reproductive function. Finally, analysis of DNA sequence polymorphism within *D. simulans* suggests the recent action of positive selection in this region of the genome, resulting in a selective sweep of sequence variation.

Molecular Evolution of the *janus* and *ocnus* Genes in the *D. melanogaster* Species Subgroup

In *D. melanogaster*, *janusA (janA), janusB* (*janB*) and *ocnus* (*ocn*) are located in a gene-dense region near the telomeric end of the right arm of chromosome 3. The genomic organization of this region is shown in Figure 1. The *janA* and *janB* transcriptional units are adjoining, with the 3' end of *janA* overlapping with the 5' end of *janB* (Fig. 1).[26] Despite this overlap, *janA* and *janB* produce separate transcripts that are under the control of independent promoters.[27] The *ocn* transcriptional unit begins approximately 250 bp downstream from the *janB* polyadenylation site, and there is no overlap between the *janB* and *ocn* transcripts (Fig. 1). Phylogenetic comparison of *janA*, *janB*, and *ocn* sequences among species of the *D. melanogaster* species subgroup, as well as *janA* and *janB* sequences from the more distantly related *D. pseudoobscura*, suggests two separate duplication events have occurred in this region of the genome (Fig. 2).[28,29]

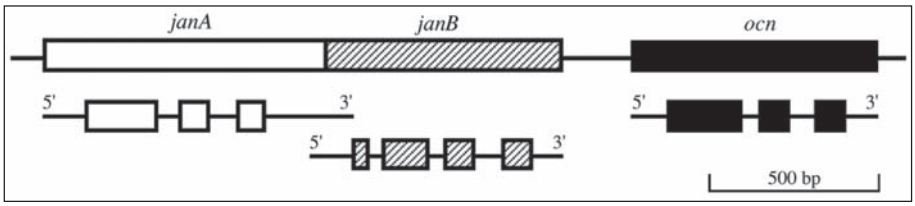

Figure 1. Organization of the *janA*, *janB*, and *ocn* genes in species of the *D. melanogaster* species subgroup. The chromosomal arrangement of the genes is shown on top, and the transcriptional units are shown below, with boxes representing protein-encoding regions and lines representing introns and untranslated regions. There is an overlap between the 3' end of the *janA* transcript and the 5' end of the *janB* transcript.

The *janB* and *ocn* genes share greater homology with each other than either does with *janA*, indicating that they are the result of a more recent duplication. Consistent with this interpretation, *janB* and *ocn* are equally divergent from *janA*. Since *janB* and *ocn* are found in all members of the *D. melanogaster* species subgroup, the duplication event that produced them must have occurred at least 10 million years ago (mya). The original duplication that produced *janA* and *janB* must predate the divergence of the *D. melanogaster* and *D. obscura* group lineages, placing this duplication event at a minimum of 25 mya. A single gene with greatest homology to *janA* was found in the *C. elegans* genome[30] suggesting that the ancestral gene was most similar in sequence to *janA*.

The evolution of these genes appears to have included increasingly restrictive function following duplication, as evidenced by their expression pattern. Experimental studies indicate that *janB* and *ocn* produce testis-specific transcripts.[26,28] In addition, translation of *janB* mRNA is restricted to the postmeiotic stages of sperm development through control elements located in the

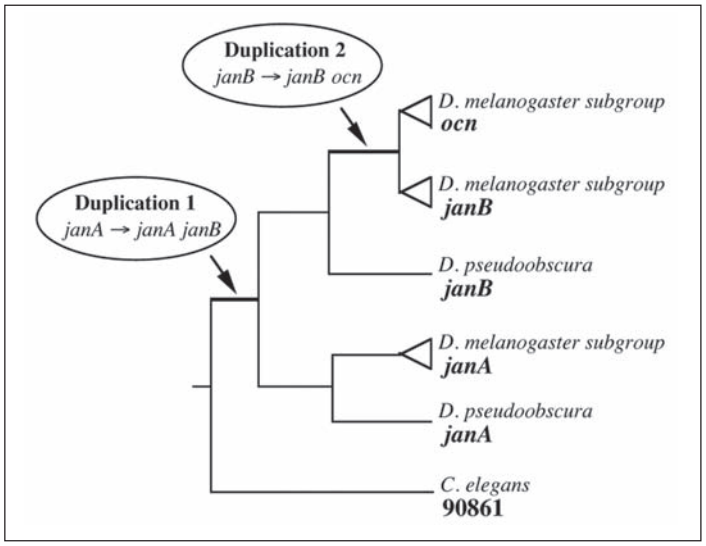

Figure 2. Gene tree of *janA*, *janB*, and ocn sequences based on protein-encoding sequences. Open triangles represent eight species of the *D. melanogaster* species subgroup. The *C. elegans* 90861 protein sequence was used to root the tree. The two gene duplication events are indicated along the branches on which they were inferred to occur.

5' UTR.[29] Conserved control elements are also found in the *ocn* 5' UTR, suggesting that it is under similar post-transcriptional regulation.[28] In contrast, *janA* produces two alternatively-spliced transcripts, one that is specific to testes and another that is found in various tissues and in both sexes.[26] The two *janA* transcripts differ in their 5' UTRs and their translation begins at different start codons, with initiation of the sperm-specific polypeptide occurring 48 bp downstream of the general initiation site.[26] As *janA* appears to be the most ancestral in sequence, the general expression pattern observed for *janA* is likely the ancestral state, with specialization to the testis-specific expression of *janB* and *ocn* occurring after duplication.

Additional support for the functional divergence of the *janA*, *janB*, and *ocn* genes comes from an analysis of *Ka*/*Ks* ratios of the three genes within the *D. melanogaster* species subgroup. If the three genes have diverged in function, then they are expected to differ in their selective constraints and potentially in their *Ka*/*Ks* ratios. To test this hypothesis, two evolutionary models were compared using a maximum likelihood approach.[28] The null model (no functional divergence) predicts that the three genes should not differ significantly from each other in their level of selective constraint, and so similar *Ka*/*Ks* ratios are expected for all three genes over all branches of the phlyogenetic tree. The alternative model predicts that following duplication each gene was subject to unique selective constraints and thus the three genes should differ in their *Ka*/*Ks* ratios. For this model, three distinct *Ka*/*Ks* ratios are expected, one each for *janA*, *janB*, and *ocn*. Maximum likelihood analysis indicates that the observed data are much more likely under the alternative model than under the null model.[28] Thus, there is strong evidence for functional and selective divergence of the three genes following duplication.

The *Ka*/*Ks* ratios for *janA*, *janB*, and *ocn* are all well below one, so there is no evidence for positive selection from the interspecific data using this strict criterion. However, all three genes have significantly higher *Ka*/*Ks* ratios than other genes (all encoding metabolic enzymes) that have been sequenced in species of the *D. melanogaster* species subgroup.[28] This observation is consistent with a general pattern of increased evolutionary rate in *Drosophila* genes with a sex-related function.[31] This increased rate of molecular evolution could have two very different explanations. One possibility is that positive selection has favored an increased fraction of amino acid replacements in reproductive genes relative to genes with other functions. This may be the result of selection on reproductive traits such as male fertility or sperm competition. The other possibility is that selective constraints are relaxed in reproductive genes, and that these genes accept more neutral amino acid changes than nonreproductive genes. Except in rare cases where the *Ka*/*Ka* ratio is significantly greater than one, such as in the accessory protein gene *Acp26Aa*,[5] it is generally not possible to distinguish between these two explanations solely through interspecific comparison of protein-encoding sequences. Additional helpful information may be gained from intraspecific studies, *i. e.* from analysis of DNA sequence polymorphism.

DNA Sequence Polymorphism in the *janus-ocnus* Region of *D. simulans*

The pattern of intraspecific DNA sequence polymorphism in the *jan-ocn* region of *D. simulans* provides evidence for the recent action of positive selection in this region of the genome.[32] A graphical representation of the polymorphic nucleotide sites in the *janA*, *janB*, and *ocn* genes of 36 *D. simulans* chromosomes sampled from a worldwide distribution is shown in (Fig. 3). In this figure, each vertical column represents a segregating site and each horizontal row represents a different chromosome. At each site, the derived variant (inferred from the *D. melanogaster* outgroup sequence) is shown in black. The unusual arrangement of variation in this sample is immediately apparent. Many alleles are identical or nearly identical in their DNA sequence, while a few are quite different. This pattern is strongest over the region containing the 3' end of *janA* and the entire *janB* gene. Here there are 16 chromosomes that are identical

Figure 3. Graphical representation of nucleotide polymorphism in the *D. simulans janA, janB*, and *ocn* genes from a worldwide sample of 36 chromosomes. Each column represents a polymorphic site, and each row represents a different chromosome. The derived variant at each site (inferred from the *D. melanogaster* sequence) is shown in black, and the ancestral in white.

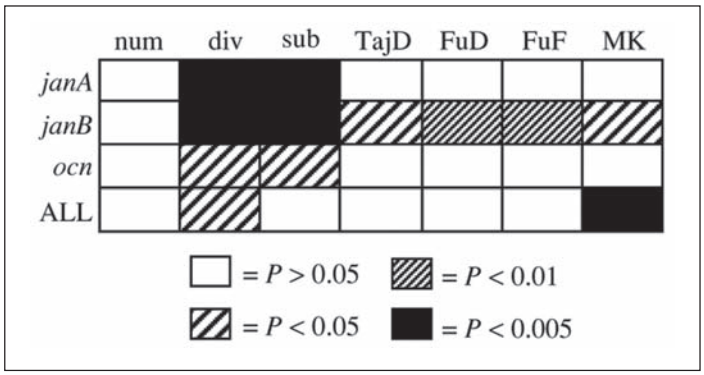

Figure 4. Results of neutrality tests applied to the *janA, janB,* and *ocn* genes. Tests were also applied to the combined data from the three genes (ALL). Column heads: num, haplotype number test;[19] div= haplotype diversity test;[19] sub= haplotype subset test;[18] Taj= Tajima's test;[17] FuD and FuF= the *D* and *F* tests of Fu and Li;[20] MK= McDonald and Kreitman's test.[25]

in their combination of variants at segregating sites (haplotype), and 9 more chromosomes that differ at only a single site. We refer to this group of 25 chromosomes as haplotype group 1 and the remaining chromosomes as haplotype group 2. Polymorphism within haplotype group 2 is in the range typically observed in *D. simulans*,[33] while haplotype group 1 shows a marked reduction in diversity. This observation, along with genealogical reconstruction of the alleles, suggests that haplotype group 2 represents a diverse collection of ancestral alleles and that haplotype group 1 represents a collection of more recently derived alleles.[32] Within haplotype group 1 there is evidence for two distinct recombination events occurring on either side of *janB*. Over the first 31 segregating sites in *janA* there are five alleles that are identical to each other, but differ from the other haplotype group 1 alleles at seven sites (Fig. 3). Similarly, there are seven *ocn* alleles that are identical to each other, but differ from the other haplotype group 1 alleles at 12 sites (Fig. 3). In both cases, the differing alleles contain many ancestral variants, indicating recombination between an allele of haplotype group 1 and an allele of haplotype group 2.

A number of statistical tests reject the neutral evolution model for the *jan-ocn* polymorphism data (Fig. 4). For example, two haplotype tests indicate that the structure of variation observed in each of the three genes differs significantly from the neutral expectation. This is due to the large number of haplotype group 1 alleles that contain very little polymorphism. The deviation is strongest for *janA* and *janB* (Fig. 4). The *janB* gene departs significantly from neutrality by several additional statistical tests, including those of Tajima[17] and Fu and Li.[20] This indicates an excess of low frequency variants and is caused primarily by the large number of singleton polymorphisms occurring within haplotype group 2. *janB* also produces a significant result for the MK test.[25] *D. simulans* and *D. melanogaster* differ at 11 synonymous sites and seven nonsynonymous sites at *janB*. Within *D. simulans*, there are 11 synonymous polymorphisms and zero nonsynonymous polymorphisms. Thus the deviation is in the direction of an excess of interspecific nonsynonymous fixations. Given that at present the amino acid sequence of *janB* in *D. simulans* appears to be under strong purifying selection, this observation suggests either relaxed selective constraint or positive selection for amino acid replacement in *janB* soon after the *D. simulans*/*D. melanogaster* split. For *janA*, *janB*, and *ocn* combined, there are 22 synonymous differences and 12 nonsynonymous differences between species. Within *D. simulans* there are 36 synonymous polymorphisms and one nonsynonymous polymorphism. This is a highly significant departure from the neutral expectation, which suggests that positive

selection may have acted not only on *janB*, but perhaps also on another gene(s) in this region since the time of the *D. melanogaster*/*D. simulans* divergence.

Distinguishing between Demographics and Selection

The age of a recently derived haplotype group can be estimated based on the *D. melanogaster*/*D. simulans* divergence and on the number of mutations observed within *D. simulans*.[34] For the *janB* sample, where two mutations are observed within 25 alleles, the age of haplotype group 1 is estimated to be ≈5000 years (95% confidence interval, 1000 - 15000 years). This indicates a rapid increase in the frequency of haplotype group 1 alleles, which currently represent 70% of a worldwide sample. Such a rapid increase in allele frequency may be the result of positive selection, but could also be explained by population demographics. For example, a recent founder event followed by rapid population expansion could also explain the observed haplotype structure. The key to distinguishing between these two possibilities is the pattern of variation at other loci on the same chromosome. Demographic factors should affect the entire chromosome in the same way, while selection should affect only particular regions of the chromosome. A survey of 19 other loci located on chromosome 3R indicates that the pattern observed in the *janA-ocn* region is highly unusual.[32,35] None of the other 19 loci shows a significant departure from neutrality by the tests described above. In addition, a survey of polymorphism in the *rp49* gene, which lies ≈ 7 kb proximal to *janA* on chromosome 3, revealed a low polymorphism haplotype group present at nearly equal frequency in *D. simulans* populations from both Europe and Africa.[34] Equal frequencies of alleles in two different populations, one presumably ancestral (Africa) and one derived (Europe), is unlikely if allele frequency is the result of founder effects.[34]

Can the pattern of variation in the *janA-ocn* region be explained by a selective sweep? Some features of the data are inconsistent with alternative explanations. First, the level of polymorphism is reduced in this region relative to other loci on the same chromosome. This reduction cannot be explained by a low mutation rate or unusually high selective constraint in this region of the genome, because there is no corresponding decrease in interspecific divergence in the *janA-ocn* region.[32] The reduced polymorphism is also unlikely to be explained by background selection, because this region of the genome does not appear to have an unusually low recombination rate.[34,36,37] Beyond the haplotype structure, another aspect of the data that is consistent with the selective sweep model is the high frequency of derived variants, which is a unique feature of genetic hitchhiking.[21,22] Although an original sample of eight *D. simulans* alleles showed a significant excess of derived variants by Fay and Wu's test,[32] the test result is not significant when applied to the larger sample of 36 alleles. This is due to the large number of low-frequency, derived variants within haplotype group 2 counteracting the high-frequency, derived variants within haplotype group 1. However, the maximum likelihood-based hitchhiking test of Kim and Stephan[38] produces a highly significant result when applied to the complete *janA-ocn* dataset (Y. Kim, personal communication), supporting the hypothesis of a recent selective sweep in this region of the genome.

A complete selective sweep is expected to eliminate variation and to drive derived variants to high frequency or fixation; however it is not expected to produce two divergent haplotype groups like those observed in the *janA-ocn* region. The observed pattern can only be explained if the sweep is incomplete. A diagram of the selective sweep model is shown in Figure 5. In this figure, a sample of 10 chromosomes is shown at three different time-points during the course of a selective sweep. The solid rectangles represent derived, neutral variants. In the first panel, these variants show an arrangement expected under neutrality. They are in low frequency and are randomly distributed among chromosomes. A new, positively-selected mutation is represented by an open rectangle. As this new variant increases in frequency in the population, linked neutral variants also increase in frequency (panel 2). As this process continues, the

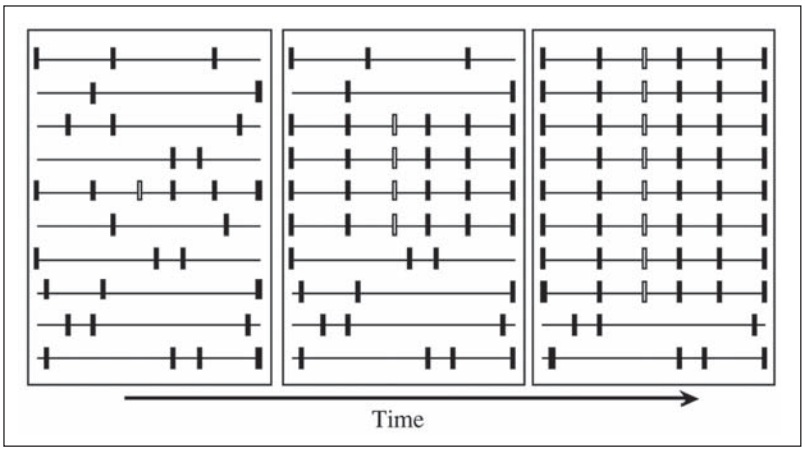

Figure 5. The genetic hitchhiking/selective sweep model. Each panel shows a sample of 10 chromosomes from a population taken at different time points. The solid boxes represent derived, neutral variants. The open boxes represent a new, positively-selected variant. As the selected variant increases in frequency, linked neutral polymorphisms "hitchhike" along. The result is a decrease in polymorphism and an increase in the frequency of derived variants. If the sweep is incomplete, two divergent haplotype groups may be present in the population.

selected variant and the linked neutral variants reach a high frequency as a single haplotype (panel 3). If this haplotype does not become fixed in the population, then some ancestral alleles may still remain present. These ancestral alleles should differ from the common haplotype at a number of sites and also differ among themselves at a level expected for a group of neutrally-evolving alleles. This pattern of variation is exactly what is observed for the *janA-ocn* region. A common haplotype group with little polymorphism and many derived variants is at high frequency, while an ancestral haplotype group with much greater variation is at lower frequency. Thus the observed data may best be explained by a selective sweep that has not gone to completion.

Why is the selective sweep incomplete? It is possible that the sweep is ongoing and that haplotype group 1 alleles will eventually become fixed in the population. This explanation seems unlikely, as the fixation of a strongly selected variant is expected to occur quite rapidly,[11] making the observation of a population in mid-sweep hardly probable. Another possibility is that the sweep is incomplete due to population subdivision within *D. simulans*. For example, haplotype group 2 alleles appear to be more frequent in African populations,[32] which are thought to be ancestral. It is possible that there is little migration of derived alleles into the ancestral populations, thus haplotype group 1 alleles do not become fixed in these populations. The problem with this explanation is that alleles from the two haplotype groups cooccur in a number of worldwide populations. Thus a population subdivision model would require a very high, nonsymmetric migration of ancestral alleles from African populations to the rest of the world for them to be present at detectable frequency. Another possibility is that there is some form of balancing selection, such as frequency dependent selection, that prevents alleles of haplotype group 1 from going to fixation. The very low level of polymorphism observed within haplotype group 1 is inconsistent with this being an old, balanced polymorphism. It indicates that if balancing selection is involved, one of the two balanced alleles must be very young and have recently been swept to its equilibrium frequency. A final possibility is that there are multiple positively-selected variants at different sites in the two haplotype groups, and fixation of a

single haplotype is delayed until a recombination event brings them together on the same chromosome. This scenario, known as the traffic model,[39] will produce a pattern of nucleotide variation similar to that seen for balancing selection until recombination brings the favored variants onto a single chromosome. The high density of genes in this region of the genome,[40] along with the general excess of interspecific amino acid replacements in the genes in the region, indicates that there is a high potential for "molecular traffic," which could result in the observed haplotype structure.

Identifying Specific Targets of Positive Selection

Identification of the particular nucleotide sites that are the target of positive selection will require a combination of population genetic and functional studies. From the intraspecific polymorphism data, the most likely location of the selected site can be inferred to be within the 3' end of *janA* or within *janB*. This region is implicated because the newly-derived haplotype is at highest frequency within this region and because there is evidence for recombination with ancestral alleles on either side (Fig. 3). If the sweep is incomplete, then the selected variant should still be segregating in the population and should be associated with haplotype group 1 alleles. This narrows the list of candidates to just a few segregating sites. Because all of the polymorphisms in this window are at synonymous or noncoding sites, any phenotypic effect must occur at the level of gene expression. Comparison of expression of genes in this region between alleles of haplotype groups 1 and 2 is therefore the first step in the attempt to elucidate phenotypic differences that may underlie genetic variation in this region. Ultimate proof of a selectively favored genetic variant requires fitness assessment of different genotypes in a controlled genetic background. However, it may be difficult to demonstrate experimentally a clear relationship between genotype and fitness in this and many other instances of putative positive selection. Among the predictable complications are the possibility that balancing selection or molecular traffic may have affected the observed haplotype pattern, the prevalence of male-female interactions affecting genes with reproductive functions, and the likelihood that selective forces operating in nature may be diminished in laboratory conditions. Hence it is gratifying and reassuring to note that statistical analysis of inter- and intraspecific DNA sequence data may have the power to detect past and ongoing natural selection in many cases where direct experimental demonstration of fitness differences is technically complicated or even impossible.

References

1. Kimura M. The neutral theory of molecular evolution. Cambridge: Cambridge University Press, 1983.
2. Endo T, Ikeo K, Gojobori T. Large-scale search for genes on which positive selection may operate. Mol Biol Evol 1996; 13:685-690.
3. Fitch WM, Bush RM, Bender CA et al. Long term trends in the evolution of H(3) HA1 human influenza type A. Proc Natl Acad Sci USA 1997; 94:7712-7718.
4. Yang Z, Nielsen R, Goldman N et al. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 2000; 155:431-449.
5. Tsaur SC, Wu CI. Positive selection and the molecular evolution of a gene of male reproduction, Acp26Aa of Drosophila. Mol Biol Evol 1997; 14:544-549.
6. Wyckoff GJ, Wang W, Wu CI. Rapid evolution of male reproductive genes in the descent of man. Nature 2000; 403:304-309.
7. Swanson WJ, Clark AG, Waldrip-Dail HM et al. Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in Drosophila. Proc Natl Acad Sci USA 2001; 98:7375-7379.
8. Yang Z, Swanson WJ. Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. Mol Biol Evol 2002; 19:49-57.
9. Maynard Smith J, Haigh J. The hitch-hiking effect of a favourable gene. Genet Res 1974; 23:23-35.

10. Kaplan NL, Hudson RR, Langley CH. The "hitchhiking effect" revisited. Genetics 1989; 123:887-899.
11. Stephan W, Wiehe THE, Lenz MW. The effect of strongly selected substitutions on neutral polymorphism: Analytical results based on diffusion theory. Theor Pop Biol 1992; 41:237-254.
12. Aguadé M, Miyashita N, Langley CH. Restriction-map variation at the zeste-tko region in natural populations of Drosophila melanogaster. Mol Biol Evol 1989; 6:123-130.
13. Stephan W, Langley CH. Molecular genetic variation in the centromeric region of the X chromosome in three Drosophila ananassae populations. I. contrasts between the vermilion and forked loci. Genetics 1989; 121:89-99.
14. Begun DJ, Aquadro CF. Molecular population genetics of the distal portion of the X chromosome in Drosophila: Evidence for genetic hitchhiking of the yellow-achaete region. Genetics 1991; 129:1147-1158.
15. Berry AJ, Ajioka JW, Kreitman M. Lack of polymorphism on the Drosophila fourth chromosome resulting from selection. Genetics 1991; 129:1111-1117.
16. Charlesworth B, Morgan MT, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. Genetics 1993; 134:1289-1303.
17. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 1989; 123:585-595.
18. Hudson RR, Bailey K, Skarecky D et al. Evidence for positive selection in the superoxide dismutase (Sod) region of Drosophila melanogaster. Genetics 1994; 136:1329-1340.
19. Depaulis F, Veuille M. Neutrality tests based on the distribution of haplotypes under an infinite-site model. Mol Biol Evol 1998; 15:1788-1790.
20. Fu YX, Li WH. Statistical tests of neutrality of mutations. Genetics 1993; 133:693-709.
21. Fay JC, Wu CI. Hitchhiking under positive Darwinian selection. Genetics 2000; 155:1405-1413.
22. Kim Y, Stephan W. Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 2002; 160:765-777.
23. Begun DJ, Aquadro CF. Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. Nature 1992; 356:519-520.
24. Hudson RR, Kreitman M, Aguadé M. A test of neutral molecular evolution based on nucleotide data. Genetics 1987; 116:153-159.
25. McDonald JH, Kreitman M. Adaptive protein evolution at the Adh locus in Drosophila. Nature 1991; 351:652-654.
26. Yanicostas C, Vincent A, Lepesant JA. Transcriptional and posttranscriptional regulation contributes to the sex-regulated expression of two sequence-related genes at the janus locus of Drosophila melanogaster. Mol Cell Biol 1989; 9:2526-2535.
27. Yanicostas C, Lepesant JA. Transcriptional and translational cis-regulatory sequences of the spermatocyte-specific Drosophila janusB gene are located in the 3' exonic region of the overlapping janusA gene. Mol Gen Genet 1990; 224:450-458.
28. Parsch J, Meiklejohn CD, Hauschteck-Jungen E et al. Molecular evolution of the ocnus and janus genes in the Drosophila melanogaster species subgroup. Mol Biol Evol 2001; 18:801-811.
29. Yanicostas C, Ferrer P, Vincent A et al. Separate cis-regulatory sequences control expression of serendipity β and janus A, two immediately adjacent Drosophila genes. Mol Gen Genet 1995; 246:549-560.
30. C. elegans sequencing consortium. Genome sequence of the nematode C. elegans: A platform for investigating biology. Science 1998; 282:2012-2018.
31. Civetta A, Singh RS. Sex-related genes, directional sexual selection, and speciation. Mol Biol Evol 1998; 15:901-909.
32. Parsch J, Meiklejohn CD, Hartl DL. Patterns of DNA sequence variation suggest the recent action of positive selection in the janus-ocnus region of Drosophila simulans. Genetics 2001; 159:647-657.
33. Moriyama EN, Powell JR. Intraspecific nuclear DNA variation in Drosophila. Mol Biol Evol 1996; 13:261-277.
34. Rozas J, Gullaud M, Blandin G et al. DNA variation at the rp49 gene region of Drosophila simulans: Evolutionary inferences from an unusual haplotype structure. Genetics 2001; 158:1147-1155.

35. Begun DJ, Whitley P. Reduced X-linked nucleotide polymorphism in Drosophila simulans. Proc Natl Acad Sci USA 2000; 97:5960-5965.
36. Hamblin MT, Aquadro CF. High nucleotide sequence variation in a region of low recombination in Drosophila simulans is consistent with the background selection model. Mol Biol Evol 1996; 13:1133-1140.
37. True JR., Mercer JM, Laurie CC. Differences in crossover frequency and distribution among three sibling species of Drosophila. Genetics 1996; 142:507-523.
38. Kim Y, Stephan W. Joint effects of genetic hitchhiking and background selection on neutral variation. Genetics 2000; 155:1415-1427.
39. Kirby DA, Stephan W. Multi-locus selection and the structure of variation at the white gene of Drosophila melanogaster. Genetics 1996; 144:635-645.
40. Adams MD, Celniker SE, Holt RA et al. The genome sequence of Drosophila melanogaster. Science 2000; 287:2185-2195.

CHAPTER 2

Rapid Evolution of Sex-Related Genes:

Sexual Conflict or Sex-Specific Adaptations?

Alberto Civetta and Rama S. Singh

Abstract

A number of recent studies have suggested that the rapid evolution of genes involved in sexual reproduction is driven by conflict between the sexes. Such genes include the ones that have a role in mating behavior, postmating gamete interactions, and fertilization (i.e., sex-related genes). However, in many cases an alternative scenario with males coadapting to female-driven changes appears as an equally likely one. Studies on the molecular evolution of sex-related genes have mainly focused on males, with few exceptions. We suggest that the combined analysis of intraspecific polymorphism and interspecies divergence will allow to make predictions on whether sex-related genes evolution is driven by conflict or coadaptation between the sexes. Such approach, made possible by rapid accumulation of DNA sequence information, will benefit from studies designed to identify male and female gene products that interact with each other during mate signaling, fertilization, and postzygotic development.

Introduction

Sexual selection captures our attention due to its effect on sexual dimorphism, evolution of exaggerated and often maladaptive traits, and its implication for the relationship between sexual conflict and fitness. The concept of sexual selection has been extended from its original meaning, which implied the evolution of secondary sexual traits that confer a mating advantage by making males better competitors against other males or more attractive to females. The extension of the concept includes not only the divergent and rapid evolution of the morphology of primary genitalia and clasping traits directly involved in mating,[1] but also that of sperm morphology[2-4] and of seminal proteins transferred in the ejaculate.[5-9] The extension comes simply from the understanding that male-to-male competition (sperm competition[10]) and/or female choice (cryptic choice[11]) are still possible even after copulation is over.

In its original form, sexual selection refers to a process leading to the development of extreme male secondary sexual traits despite of their potentially detrimental effect on survival. While natural selection draws our attention to differences in viability, sexual selection focuses on differences in mating success. It is usually entertained that males increase their overall fitness at the expense of survival by increasing their chances to be chosen for mating. Females increase their fitness in an indirect way by choosing males that will provide them with successful male progeny (thus propagating the female's genes). The result is a constant selection of males with the appropriate signals until the male traits become so elaborate or so extreme that

Selective Sweep, edited by Dmitry Nurminsky. ©2005 Eurekah.com and Kluwer Academic/Plenum Publishers.

the balance between survival and reproductive advantage is upset by the high cost of the trait. Although fitness is presented in terms of separate (male or female) sexes, the end result is an overall increase in fitness at the population level.

However, a side consequence of thinking in terms of distinct sexes is a tendency to appraise fitness of males and females separately. When such dissociation is made, female's fitness is usually considered in terms of her overall reproductive advantage rather than in terms of her viability. In recent years, the idea that the viability aspect of female's fitness could be impaired by the evolution of male traits that increase his mating success has gained support. The result is the concept of conflict, or arm races, between the sexes—a constant struggle in which male traits evolve to extremes to secure mating at the expense of being detrimental to their partners.

While it has become customary to calculate fitness for each sex separately, it is obvious that the population fitness depends on the interaction and coadaptation of both sexes. Separate sexes cannot maximize their fitness indefinitely, and not all reproductive traits are subject to sexual conflict. Sex-specific signals to the members of the conspecific opposite sex, complementarity of genitalic structures, and egg-sperm interactions are just a few examples of traits that would be subject to coadaptation rather than to conflict. Clasping structures, mating frequency, and sperm competition are the obvious examples of conflict traits. However, in many cases it is not clear whether the trait is likely to have evolved under male-female coadaptation or sexual conflict scenario. We propose that such cases could be resolved by analysis of the dynamics of genetic polymorphism and divergence for candidate genes under the contrasting scenarios of sexual conflict *vs*. sexual coadaptation.

Are Males Conflicting with Females or Coadapting to Them?

The use of the "conflict" or "war" term is not restricted to the field of sexual selection and evolution. It is a widespread and attractive term that is widely applied to political, sociological, and health related issues (the war against cancer, the war on drugs, etc).

There are clear examples of the harm potentially inflicted by males to females during courtship and mating. At the mating level, it is usually forgotten that such harm is not necessarily beneficial to the male. For example, it has been recently shown that even though sexual selection may favor male bowerbirds that display intensively and therefore appear more attractive, a tradeoff is in place because excessive intensity is perceived as aggressive behavior threatening to females; such potentially harmful performance leads to an end of courtship.[12] This is a very important idea in terms of the potential effect of conflict between the sexes because during courtship behaviors should favor communication that will benefit both sexes. In a situation where males were allowed to overcome male-females tradeoffs in terms of access to mating, males with higher sexual activity (access to females) showed a diminished antibacterial immune reaction, suggesting a cost of male sexual activity that is irrelevant to the female's response.[13]

In a recent paper, Arnqvist and Rowe demonstrated a constant coadaptation between the sexes in water striders, with deviations from such male-female coadaptations leading to rapid evolution.[14] In the species analyzed, males have evolved clasping genitalia to grasp females, and females have evolved counteradaptations to resist grasping, suggesting an arm-race between the sexes. However, it is not usually easy to establish whether male-female coevolution results as a consequence of conflict or of mutual benefit between the sexes. In species where there is male-to-male competition, developing of traits that facilitate access to females or improve sperm competitiveness may have side effects deleterious to females.[14,15] Such pattern of sexual conflict has been proposed to promote speciation, because polyandrous species show a higher speciation rate.[16] Another possible explanation of high speciation rate is coadaptation leading to fine-tuning of male and female reproductive systems, as seems to be the case for beetles where sperm of males from the same population as females outcompetes sperm of male from an allopatric population.[17]

Among *Drosophila* species, the evolution of long sperm is correlated with longer sperm storage organs in females.[18] Experiments with selection for divergent sperm length or size of female sperm storage organs in *D. melanogaster* revealed that male-female coadaptation is responsible for such correlations.[19] *Drosophila* females' lifespan is affected due to male seminal fluids transferred in the ejaculate,[20] with the effect being more drastic when the male is a strong sperm competitor.[15] An interesting observation is that despite the cost of mating, mated females live longer than virgins.[15] Similar results have been shown for the Meditaeranean fruit fly *Ceratitis capitata* where virgin females experience lower mortality than mated females only during the first 20 days after eclosion.[21] It appears that there is a compensation to the cost on viability that females have to pay shortly after mating, as mated females live longer than virgins.[15,21]

Sexual conflict has been implicated in rapid evolution of genes with a role in fertilization, but obviously it does not represent a single driving force behind the phenomenon. In other words, rapid evolution of a gene involved in fertilization does not necessarily imply a sexual conflict. In this chapter we first examine what we have learned from studies on rapid molecular evolution of immune system genes, where there is a clear arm-race between the infective agent and the infected organism. We then explore the pattern of rapid molecular evolution of genes with a role in mate recognition, sperm competition and/or fertilization (i.e., sex-related genes). Finally, we make predictions of what would be commonly observed for genes rapidly evolving under conflict *versus* those simply involved in a rapid coadaptation between males and females.

Conflict Scenarios: Balancing Selection in Rapidly Evolving Immune System Genes

One of the best-understood examples of adaptive evolution is that of the Major Histocompatibility Complex (MHC) in vertebrates. MHC proteins are expressed on the cell surface where they present small peptides to cytotoxic T lymphocytes or helper T cells. The efficiency of the immune response to battle a wide variety of infections depends on the ability of the MHC system to present a wide variety of peptides that can be recognized as foreign. The resolution of the molecular structure of MHC molecules[22,23] was crucial for understanding of the role played by selection in the evolution of these proteins. The peptide binding region of MHC molecules became a logical target for the prediction of an elevated polymorphism that will allow a wide variety of alleles capable of recognizing and presenting multiple antigens. Indeed, the peptide binding region has the highest proportion of nonsynonymous to synonymous polymorphic sites within the protein,[24,25] indicative of balancing selection. The level of polymorphism at peptide binding regions apparently depends on the level of variability of the peptides that it binds.[26] Therefore, parasites represent the major selective force that drives evolution of the protein-binding region of MHC molecules.

On the other side of the barricades, there are parasites that try to escape recognition by the immune system. For example, in *Plasmodium falciparum*, an unusually high proportion of nonsynonymous substitutions was found in the region of the circumsporozoite protein (CSP) which provides the peptides recognized by the MHC molecules.[27] An additional intriguing twist to the arm-race scenario in the case of malaria is introduced by cross-reactivity between the epitopes of *Plasmodium falciparum* and the cytotoxic T-lymphocytes (CTL) epitopes, which results in selective pressure on human leukocyte proteins.[28] During HIV infection, the host antibody response can act as a strong selective agent targeting surface proteins of the virus. Sequence analyses of a series of hypervariable regions in these proteins suggest that amino acid polymorphisms are favored by positive selection, but that different lineages of the virus show different responses depending on the nature of the host environment.[29] Again, an arm-race between infectious agents and the host leads to the rapid accumulation of polymorphisms that can help escape recognition by the host immune system.[30,31]

Other recognition systems, such as self-incompatibility genes in plants, and pheromone-receptor recognition in fungi, also show evidence for balancing selection by preservation of multiple alleles.[32-34] The common pattern is high polymorphism within species combined with rapid divergence between species.[35]

External Fertilization: How Are Marine Invertebrate Sperm Surface Proteins Different from Immune System Genes?

In sea urchins, the sperm surface protein bindin mediates binding to the egg. The protein is not only highly divergent between species but also highly polymorphic within species.[36-38] The high polymorphism within species could suggest a situation of sexual conflict with male bindin proteins being in a state of male-to-male competition for the eggs. In a sexual conflict scenario where males control the results of fertilization, the bindin proteins for which females have not yet coevolved proper defense would be more successful. However, Palumbi[39] showed that sea urchins could be grouped into different clades according to the sequence of their bindin alleles, and that males are more efficient at fertilizing eggs of the same genotype, suggesting female cryptic choice scenario as opposed to sexual conflict. Male-to-female coadaptation within the population, where different bindin alleles work better with akin partners, apparently have maintained high polymorphism and the system seems to be female driven.

In another marine invertebrate, abalone, both the sperm surface protein (lysin) and its receptor on the egg (VERL) have been identified. Swanson and Vacquier[40] have shown that the lysin-binding motif of VERL is a large glycoprotein with 22 tandem repeats which shows rapid divergence, but no sign of positive selection. They demonstrated that concerted evolution of the VERL repeats leads to species-specific phylogenetic grouping, and proposed that redundancy of its tandem repeat structure reduces selective pressures on the receptor. The spread of accumulated mutations within VERL by unequal crossing-over and gene conversion imposes new selective pressures on lysin, so that species-specific fertilization evolves as a consequence of males having to adapt to a rapidly neutrally evolving sperm receptor on the egg surface.[40,41]

The examples coming from studies on marine invertebrates with an external fertilization system show a situation in which female sperm receptors are not only essential for proper fertilization, but also are the source of selective pressure on male sperm proteins that are constantly adapting to the female-driven changes. The marine invertebrate system clearly shows the need for a proper understanding of who are the key players in sperm-egg interactions during fertilization, before we can make predictions as to whether rapid evolution of male reproductive proteins driven by positive selection is a consequence of sexual conflict or of male-to-female coadaptation.

Internal Fertilization: Lessons from *Drosophila*

Drosophila males produce secretions in a pair of accessory glands (called also paragonia) and transfer these secretions in their ejaculate. The paragonia represent a component of the male's internal genitalia. The potential effect of accessory gland proteins on the female physiology and behavior has been suggested in different experimental settings. Accessory gland proteins transferred during copulation have been found to be responsible for triggering a series of postcopulatory female responses, such as increased ovulation and delayed remating.[42]

Several studies also suggested that accessory gland secretions are capable of affecting female's longevity after mating. When *Drosophila* males lacking the accessory gland secretions are compared to intact males, only the males that transfer accessory gland secretions during copulation are able to reduce average longevity of female.[20] The specific factors that are responsible for this effect remain elusive. However, the most likely candidate is *Acp62F*, an accessory gland protein capable of entering the haemolymph of females after mating,[43] and the only accessory gland product proven to be toxic to females and males when ectopically expressed.[44]

Why would males transfer protein(s) that inflict a damage to females by reducing their viability? One likely possibility is that the transfer of such proteins provides a benefit in terms of male reproductive ability. We know that there is a significant and negative correlation between sperm competitiveness and female viability shortly after mating,[15] and it appears that mating provides a long-term survival advantage for females.[15,21] An antagonistic arm race between the sexes may start as a consequence of a male toxicity, with females evolving responses to escape the harmful effect of male's ejaculate and with males readjusting to the new female environment. Specific mechanisms on the female side could include a rapid evolution of female receptors to escape effects of the male proteins or, perhaps, developing of female reproductive gland secretions that can neutralize the male's accessory gland secretions. A likely scenario is the one in which males have evolved mechanisms that harm females after mating, because it provides an immediate advantage in terms of male-to-male competition. Such situation is tolerated, because females acquired a mating-induced long-term survival advantage. Therefore, the connection between competitiveness of male sperm and cost of mating in terms of female longevity seems to be more complex than a simple toxic effect inflicted by seminal fluids.

Arm-Race vs. Sex-Specific Coadaptation: Test of Hypotheses

Female Response: Are There Female Counterparts for Male Proteins?

Very little is known about the pattern of molecular evolution of female reproductive proteins. Civetta and Singh[45] showed that proteins expressed in *Drosophila* ovary tissue are as highly divergent between species as are male reproductive proteins. More recently, it has been shown that zona pellucida proteins in mammalian eggs display signs of rapid divergence shaped by positive selection.[46]

How helpful is the information on the evolutionary patterns of female reproductive tissue proteins in deciding between the conflict *versus* the coadaptation scenarios? While the conflict scenario is male driven, the lack of conflict implies that females are choosing and males are coadapting to the new requirements introduced by females. It is possible to test conflict vs. coadaptation by examining the combined pattern of molecular evolution of male and female proteins that interact at various stages of gamete recognition and fertilization. Under the coadaptation hypothesis, the expectation is that female proteins will show a higher proportion of differences between populations than the male counterparts, at least in early stages of differentiation between populations. However, the pattern might not hold in a case where the female component (e.g., sperm receptor, reproductive gland proteins) is duplicated and therefore there is no one-to-one relationship between the male and the female counterparts, as in the case of the lysin/VERL system.[40,47]

A conflict scenario will predict males driving divergence between populations and therefore male proteins having higher levels of polymorphism. A major limitation to the test is the lack of identified female reproductive proteins that interact with well-characterized male counterparts such as the accessory gland proteins in *Drosophila*. A general picture of the situation in *Drosophila* can be inferred from our study on protein divergence between pairs of species from the *virilis* group.[45] This is an interesting group, since it offers pairs of species with different levels of postmating prezygotic isolation, from the pairs capable of producing viable and fertile offspring to the pairs producing no viable/fertile progeny. Comparisons between species belonging to the *virilis* phylad, where flies are capable of interspecific hybridization yielding fertile progeny, show that the level of protein divergence found in ovaries is higher than that in testes for some species pairs. This observation suggests that females are driving the direction of evolution of male traits, implying the coadaptation scenario. However, larger samples and a more detailed analysis of reproductive isolation in the *virilis* group species are needed for a convincing claim.

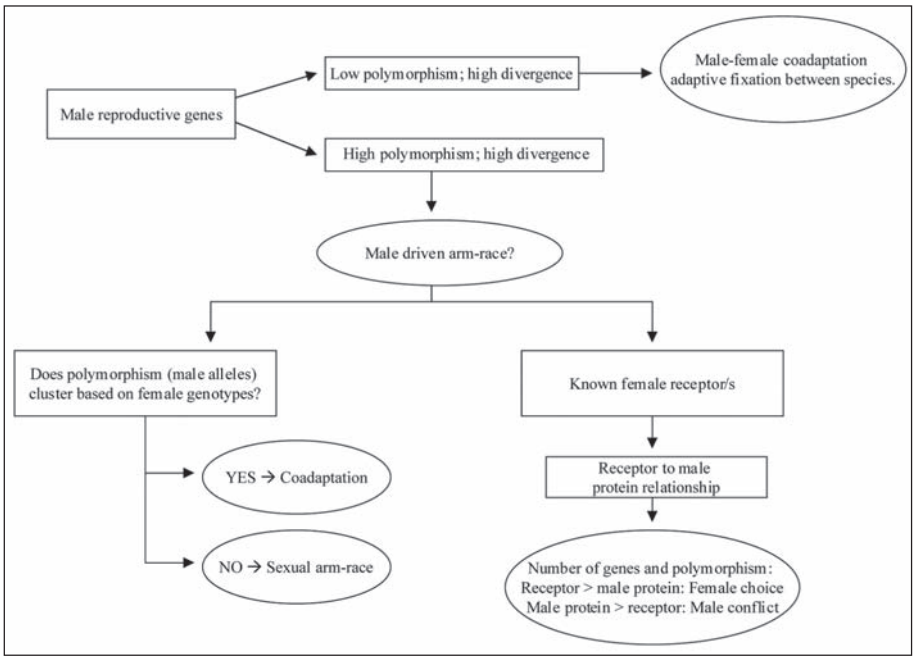

Figure 1. Predictive patterns of genetic polymorphism and interspecific divergence for genes evolving under sexual conflict or sexual coadaptation.

Another more recent study in mammals analyzed the pattern of molecular evolution of three zona pellucida proteins that are either known or suspected to play a key role during sperm recognition and binding. All three proteins showed signs of positive selection, and specific residues that may be involved in species-specific gamete interactions were identified.[46] However, to formally test whether such pattern of positive selection is male (conflict) or female (cryptic choice) driven, we need a simultaneous analysis of the pattern of evolution of the interacting sperm proteins, including the amount and pattern of divergence and phylogenetic clustering of male and female alleles (Fig. 1).

Male Reproductive Proteins: Patterns of Polymorphism and Divergence

Given the current scarcity of data on molecular evolution of female reproductive genes, assumptions on the mode of evolution of sex-related genes are often based on analysis of just the male reproductive genes. While this approach lacks the meticulousness of the aforementioned simultaneous analysis of the male and female counterparts, it is valid for discriminating between the two possible rationales for the rapid divergence of reproductive genes, i.e., between the conflict and the coadaptation scenarios. If conflict between the sexes is in place, the expectation for the patterns of molecular evolution should be similar to that of a parasite-host situation, where the alleles of female genes which confer resistance to male harmful proteins are beneficial. We expect that such balancing selection would promote an elevated proportion of replacement polymorphisms and would result in coexistence of multiple ancient alleles (Fig. 1). The effect of selection could be localized at specific sites or regions within the genes, or could be more widespread as a consequence of a hitchhiking effect (selective sweep). The strength of the sweep depends on the level of intragenic recombination. The expectation in terms of the molecular evolutionary pattern between species depends on the intensity of selection, but it is

possible that we might see polymorphic sites shared between species, as well as fixed differences.

An alternative scenario is the one of sex-specific coadaptation, where male and female genes involved in reproduction are under purifying selection within species, showing low polymorphism (in particular, nonsynonymous polymorphism). In regions of low recombination, the removal of deleterious mutant alleles will strongly reduce polymorphism according to the background selection model.[48] If adaptive selective pressures in different species differentially fix nonsynonymous variants in these genes, an elevated proportion of nosynonymous substitutions is expected between species. Such situation is anticipated if adaptive differentiation at sex-related genes is driving speciation. The prediction is therefore one of low polymorphism, in particular nonsynonymous polymorphism, within species but elevated proportion of nonsynonymous changes between species (Fig. 1).

Conclusion

We expect natural selection to optimize sex-related traits with respect to their effects on reproduction and fitness. However, sexual inequality created by differences in allocation of resources, and competition within or between the sexes, can lead to evolution of exaggerated and often maladaptive traits. The concept of sexual selection has been extended to include both pre copulation and post-copulation interactions between males and females or between their gametes, respectively. The traits covered under this broad definition of sexual selection[35] encompass those that have evolved by sexual coadaptation as well as by sexual conflict. In this chapter we compared the evolutionary dynamics of the genes involved in pathogen/host interactions and of the sexual reproduction genes, and made predictions on the patterns of genetic polymorphism and divergence for genes experiencing sexual coadaptation *vs*. sexual conflict. The predictions provide a tool to deduce reasons for the rapid evolution of individual sex-related genes. This approach will utilize rapidly accumulating whole-genome DNA sequence data, thus providing a wealth of information on the evolutionary dynamics of sex-related genes.

References

1. Eberhard WG. Sexual selection and animal genitalia. Cambridge: Harvard University Press, 1985.
2. Pitnick S, Markow TA. Male gametic strategies: Sperm size, testes size, and the allocation of ejaculate among successive mates by the sperm-limited fly Drosophila pachea and its relatives. Am Nat 1994; 143:785-819.
3. Joly D, Bressac C, Lachaise D. Disentangling giant sperm. Nature 1995; 377:202.
4. Karr TL, Pitnick S. The ins and outs of fertilization. Nature 1996; 379:405-406.
5. Coulthart MB, Singh RS. High level of divergence of male-reproductive-tract proteins, between Drosophila melanogaster and its sibling species, D. simulans. Mol Biol Evol 1988; 5:182-191.
6. Aguadé M, Miyashita N, Langley CH. Polymorphism and divergence in the Mst26A male accessory gland gene region in Drosophila. Genetics 1992; 132:755-770.
7. Tsaur S-C, Wu C-I. Positive selection and the molecular evolution of a gene of male reproduction, Acp26Aa of Drosophila. Mol Biol Evol 1997; 14:544-549.
8. Aguadé M. Different forces drive the evolution of the Acp26Aa and Acp26Ab accessory gland genes in the Drosophila melanogaster species complex. Genetics 1998; 150:1079-1089.
9. Aguadé M. Positive selection drives the evolution of the Acp29AB accessory gland protein in Drosophila. Genetics 1999; 152:543-551.
10. Parker GA. Sperm competition and its evolutionary consequences in insects. Biological Reviews 1970; 45:525-567.
11. Eberhard WG. Female control: Sexual selection by cryptic female choice. Princeton: Princeton University Press, 1996.
12. Patricelli GL, Uy JAC, Walsh G et al. Male displays adjusted to female's response. Nature 2002; 415:279-280.

13. McKean KA, Nunney L. Increased sexual activity reduces male immune function in Drosophila melanogaster. Proc Natl Acad Sci USA 2001; 98:7904-7909.
14. Arnqvist G, Rowe L. Antagonistic coevolution between the sexes in a group of insects. Nature 2002; 415:787-789.
15. Civetta A, Clark AG. Correlated effects of sperm competition and post-mating female mortality. Proc Natl Acad Sci USA 2000; 97:13162-13165.
16. Arnqvist G, Edvardsson M, Friberg U et al. Sexual conflict promotes speciation in insects. Proc Natl Acad Sci USA 2000; 97:10460-10464.
17. Brown DV, Eady PE. The functional incompatibility between the fertilization systems of two allopatric populations of Callosobruchus maculatus (Coleoptera: Bruchidae). Evolution 2001; 55:2257-2262.
18. Pitnick S, Markow TA, Spicer GS. Evolution of multiple kinds of female sperm-storage organs in Drosophila. Evolution 1999; 53:1804-1822.
19. Miller GT, Pitnick S. Sperm-female coevolution in Drosophila. Science 2002; 298:1230-1233.
20. Chapman T, Liddle LF, Kalb JM et al. Cost of mating in Drosophila melanogaster females is mediated by male sccessory gland products. Nature 1995; 373:241-244.
21. Carey JR, Liedo P, Harshman L et al. A mortality cost of virginity at older ages in female Mediterranean fruit flies. Exp Gerontol 2002; 37:507-512.
22. Lawlor DA, Zemmour J, Enis PD et al. Evolution of class-I MHC genes and proteins: From natural selection to thymic selection. Annu Rev Immunol 1990; 8:23-63.
23. Jardetzky TS, Brown JH, Gorga JC et al. Three-dimensional structure of a human class II histocompatibility molecule complexed with superantigen. Nature 1994; 368:711-718.
24. Hughes AL, Nei M. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 1988; 335:167-70.
25. Hughes AL, Hughes MK, Howell CY et al. Natural selection at the class II major histocompatibility complex loci of mammals. Philos Trans R Soc Lond B Biol Sci 1994; 346:359-367.
26. Hughes AL, Hughes MK. Natural selection on the peptide-binding regions of major histocompatibility complex molecules. Immunogenetics 1995; 42:233-243.
27. Hughes AL, Hughes MK. Natural selection on Plasmodium surface proteins. Mol Biochem Parasitol 1995; 71:99-113.
28. Gilbert SC, Plebanski M, Gupta S et al. Association of malaria parasite population structure, HLA, and immunological antagonism. Science 1998; 279:1173-1177.
29. Seibert SA, Howell CY, Hughes MK et al. Natural selection on the gag, pol, and env genes of human immunodeficiency virus 1 (HIV-1). Mol Biol Evol 1995; 12:803-813.
30. Klenerman P, Rowland-Jones S, McAdam S et al. Cytotoxic T-cell activity antagonized by naturally occurring HIV-1 Gag variants. Nature 1994; 369:403-407.
31. Price DA, Goulder PJR, Klenerman P et al. Positive selection of HIV-1 cytotoxic T lymphocyte escape variants during primary infection. Proc Natl Acad Sci USA 1997; 94:1890-1895.
32. Uyenoyama MK. Genealogical structure among alleles regulating self-incompatibility in natural populations of flowering plants. Genetics 1997; 147:1389-1400.
33. Saupe SJ, Glass NL. Allelic specificity at the het-c heterokaryon incompatibility locus of Neurospora crassa is determined by a highly variable domain. Genetics 1997; 146:1299-1309.
34. Vaillancourt LJ, Raudaskowski M, Spetch CA et al. Multiple genes encoding pheromones and a pheromone receptor define the Bβ1 mating-type specificity in Schizophyllum commune. Genetics 1997; 146:541-551.
35. Civetta A, Singh RS. Broad-sense sexual selection, sex gene pool evolution, and speciation. Genome 1999; 42:1033-1041.
36. Metz EC, Palumbi SR. Positive selection and sequence rearrangements generate extensive polymorphism in the gamete recognition protein bindin. Mol Biol Evol 1996; 13:397-406.
37. Metz EC, Gomez-Gutierrez G, Vacquier VD. Mitochondrial DNA and bindin gene sequence evolution among allopatric species of the sea urchin genus Arbacia. Mol Biol Evol 1998; 15:185-195.
38. Metz EC, Robles-Sikisaka R, Vacquier VD. Nonsynonymous substitution in abalone sperm fertilization genes exceeds substitution in introns and mitochondrial DNA. Proc Natl Acad Sci USA 1998; 95:10676-10681.

39. Palumbi SR. All males are not created equal: Fertility differences depend on gamete recognition polymorphisms in sea urchins. Proc Natl Acad Sci USA 1999; 96:12632-12637.
40. Swanson WJ, Vacquier VD. Concerted evolution in an egg receptor for a rapidly evolving abalone sperm protein. Science 1998; 281:710-712.
41. Swanson WJ, Vacquier VD. The rapid evolution of reproductive proteins. Nat Rev Genet 2002; 3:137-144.
42. Wolfner MF. The gifts that keep on giving: Physiological functions and evolutionary dynamics of male seminal proteins in Drosophila. Heredity 2002; 88:85-93.
43. Lung O, Wolfner MF. Identification and characterization of the major Drosophila melanogaster mating plug protein. Insect Biochem Mol Biol 1999; 29:1043-1052.
44. Lung O, Tram U, Finnerty CM et al. The Drosophila melanogaster seminal fluid protein Acp62F is a protease inhibitor that is toxic upon ectopic expression. Genetics 2002; 160:211-224.
45. Civetta A, Singh RS. High divergence of reproductive tract proteins and their association with postzygotic reproductive isolation in Drosophila melanogaster and Drosophila virilis group species. J Mol Evol 1995; 41:1085-1095.
46. Swanson WJ, Yang Z, Wolfner MF et al. Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. Proc Natl Acad Sci USA 2001; 98:2509-2514.
47. Nei M, Zhang J. Molecular origin of species. Science 1998; 282:1428-1429.
48. Chalesworth B, Morgan MT, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. Genetics 1993; 134:1289-1303.

CHAPTER 3

Selective Sweep in the Evolution of a New Sperm-Specific Gene in *Drosophila*

Rob J. Kulathinal, Stanley A. Sawyer, Carlos D. Bustamante, Dmitry Nurminsky, Rita Ponce, José M. Ranz and Daniel L. Hartl

Abstract

The *Sdic* gene cluster at the base of the X-chromosome is unique to the lineage of *Drosophila melanogaster*. The repeating unit in the cluster was formed from a duplication and fusion of the genes, *AnnX* and *Cdic*, which juxtaposed the 3' untranslated region of *AnnX* to the third intron of *Cdic*. *AnnX* encodes Annexin 10 and Cdic encodes a cytoplasmic dynein intermediate chain. The 3' untranslated region of *AnnX* contains two promoter elements, including a testis-specific element, and *Cdic* intron 3 contains a third promoter element; together these elements result in testis-specific transcription of *Sdic*. The Sdic protein features a novel amino terminus derived in part from *Cdic* intron 3 which contains motifs similar to those in axonemal dyneins. It has been demonstrated that the Sdic protein becomes incorporated into the tails of mature sperm. The evolution of the *Sdic* cluster required several deletions, at least one insertion, at least eleven nucleotide substitutions, and an estimated tenfold tandem duplication, all of which took place in the 1–3 million years since the divergence of *D. melanogaster* from *D. simulans*. Evidence for the ongoing evolution of *Sdic* including a recent selective sweep is found in the low levels of polymorphism across neighboring genes in the region, a large number of fixed amino acid replacements relative to fixed synonymous nucleotide substitutions, and a frequency spectrum of polymorphic nucleotides skewed toward rare variants. The analysis of polymorphism and divergence in the *Sdic* region, however, is complicated by the possible effects of background selection caused by deleterious new mutations, owing to the reduced amount of recombination in the region associated with its proximity to centromeric heterochromatin. We present the rapid evolution of this novel gene as a fascinating example of male-driven evolution incurred by recurrent selective sweeps.

Introduction

Recent analyses of amino acid polymorphisms within species and differences between species of *Drosophila* have provided evidence that amino acid replacements are frequently driven by positive selection.[13,21,51] In all three analyses, the principal conclusion rests primarily on the observation that the ratio of amino acid replacements to synonymous nucleotide substitutions between species is greater than the ratio of amino acid polymorphisms to synonymous nucleotide polymorphisms within species.[33,48] From their analysis of polymorphism and divergence in *D. simulans* and *D. yakuba*,[51] Smith and Eyre-Walker (2002) deduce that about 45% of the

Selective Sweep, edited by Dmitry Nurminsky. ©2005 Eurekah.com and Kluwer Academic/Plenum Publishers.

amino acid replacements between these species have been driven by positive selection. Their data suggest that these species have undergone one amino acid replacement every 20 years (~200 generations), or about 600,000 substitutions altogether, of which 270,000 were driven by selection. Fay et al[21] have carried out a similar analysis of data from 45 genes in *D. melanogaster* and *D. simulans* and have come to a somewhat different conclusion. Although they also noted strong evidence for positive selection in the data as a whole, they attributed most of the positive selection to 11 genes (*Acp26Aa*, *Acp29Ab*, *anon1A3*, *anon1E9*, *anon1G5*, *ci*, *est-6*, *Ref2P*, *Rel*, *tra* and *Zw*) and regarded the remaining 34 genes as evolving essentially neutrally with respect to amino acid replacements.

Bustamante et al[13] have carried out a hierarchical Bayesian analysis of polymorphism and divergence data, using a set of 34 genes in *D. melanogaster* and *D. simulans*, partly overlapping the set of genes analyzed by Fay and colleagues.[21] We found the Bayesian approach appealing because the data are analyzed in the aggregate to estimate the average selection coefficient of each gene individually, and each estimate has an accompanying 95% credible interval, which is the Bayesian analog of the 95% confidence interval. The credible intervals emerge naturally because the Bayesian analysis is implemented by a Markov chain Monte Carlo stochastic process whose stationary distribution coincides with the posterior distribution of the parameters conditional on the data (Gilks et al, 1996).

The Bayesian analysis on the *Drosophila* data yields average scaled selection coefficients, $N_e s$, ranging from -1.12 to +4.12, where N_e is the haploid effective population size and s is the conventional selection coefficient. Among the 34 estimates, 32 are positive, again suggesting an important role for positive selection. Included among the most strongly positively selected genes, whose 95% credible interval does not overlap zero, are *Acp26Aa*, *Acp29Ab*, *anon1A3*, *anon1E9*, *ci* and *Zw*, which are found on the list of eleven rapidly evolving genes to which Fay et al[21] attribute most of the positive selection. Three genes in their list (*anon1G5*, *est-6* and *Ref2P*) are not among the most strongly positively selected genes in the Bayesian analysis, however, but are intermixed among the others. Hence, the Bayesian analysis supports that of Fay et al[21] but not completely.

The Bayesian analysis also supports that of Smith and EyreWalker,[51] but again not completely. Considering the 95% credible intervals across all genes, about 80% of the total span of the credible intervals is positive. This is much larger than the 45% positively selected amino acid replacements estimated in their study.[51] However, 57% of the total span of the credible intervals has $N_e s > 1$ and 49% has $N_e s > 2$; likewise 65% of the mean values of $N_e s$ are greater than one and 38% are greater than two. These proportions of positively selected amino acid replacements can be reconciled with those of Smith and EyreWalker[51] if their method identifies amino acid replacements as positively selected provided that $N_e s > \sim 2$.

Details of the analyses aside, there seem to be a significant number of amino acid replacements that are driven by positive selection. As judged by the Bayesian analysis, however, the intensity of selection is relatively small. Across all genes, the average value of $N_e s$ equals 1.5. This intensity of selection is sufficiently weak that genetically linked neutral polymorphisms would hardly be affected unless the linkage is very tight.[57]

Yet there is also considerable evidence for "selective sweeps" which describes positive selection of a certain magnitude affecting linked neutral variation.[31] Its presence is revealed by nonneutral haplotype frequencies, typically as an excess of rare alleles across a region of the genome or as an excess in the frequency of a single haplotype. Although the interpretation of such observations is potentially complicated by demographic factors such as population subdivision, changes in population size, or founder effects, examples of apparent selective sweeps in *D. melanogaster* include regions containing the genes, *Sod*,[26] *white*,[28,29] *Suppressor of Hairless*[20] and *Fbp2*.[10] In *D. simulans*, they include regions containing the genes, *Pgd*,[9] *runt*,[30] *Zw* and *vermilion*,[23] and *ocnus*.[42]

In this paper, we summarize evidence for one or more selective sweeps in the region of a newly evolved gene found on the X-chromosome of *D. melanogaster*. The gene, denoted *Sdic*, encodes the intermediate chain for an axonemal dynein; it is expressed specifically in the testes and its novel protein is incorporated into the mature sperm tails.[37] The novel gene is found only in *D. melanogaster* and not in any of its sibling species, including *D. simulans*.[37] We first examine what is known about the origin and genetic structure of *Sdic*, examine the evidence for one or more selective sweeps, describe the results of a hierarchical Bayesian analysis of polymorphism and divergence in the *Sdic* region and briefly discuss *Sdic*'s rapid divergence in the more general context of the faster evolution of male-specific genes. The emphasis in this paper is on the evidence for selective sweeps. Further details about the origin and molecular structure of *Sdic* can be found in reference 45.

The Origin of *Sdic*

The *Sdic* gene was discovered through an anomalous cDNA sequence recovered in a study of alternative splicing of cytoplasmic dynein intermediate-chain transcripts.[36] Dynein intermediate chains are one component of the multisubunit dynein complex whose function in the cytoplasm is to act as a minus end-directed microtubule motor.[27,43] In *Drosophila*, the multiple forms of the dynein intermediate chains are created by alternative splicing of the transcript of a single-copy gene, denoted *Cdic*, located in polytene chromosome region 19A near the base of the X-chromosome.[36]

The anomalous intermediate chain cDNA was unusual in that the apparent amino end of the coding sequence was missing two conserved amino-terminal domains necessary for interacting with proteins that help attach the dynein complex to its cytoplasmic targets. Instead, the amino-terminal end of the protein had a novel sequence resembling axonemal dynein intermediate chains.[37] The intermediate chains of the axonemal dyneins are localized at the base of dynein complex and are thought to bind directly to the A-microtubule.[43] In a genomic clone containing the coding sequence for the anomalous cDNA, the region upstream from the transcription start site was a sequence closely resembling the single-copy gene, *Annexin X* (denoted *AnnX*), which encodes one of a large family of proteins that bind to phospholipids in a calcium-dependent manner and appears to have a wide variety of functions.[7,22] It soon became apparent that, in *D. melanogaster*, both *Cdic* and *AnnX* had been duplicated, and that the anomalous cDNA resulted from a gene fusion that is expressed specifically in the testes and that encodes a putative cytoplasmic dynein intermediate chain that becomes incorporated into the axoneme of the tail of the mature sperm.[37]

In the genome of *D. simulans* and other sibling species of *D. melanogaster*, the orthologs of *Cdic* and *AnnX* are situated in the order, Telomere · · · –*AnnX*–*Cdic*– · · · Centromere, and transcription of each gene takes place from right to left. In the origin of *Sdic*, it is clear that there was a duplication of the region including *AnnX* and *Cdic*, leading to the structure, Telomere · · · –*AnnX*–*Cdic*–*AnnX*–*Cdic*– · · · Centromere. A series of deletions fused the middle two genes in such a way that intron 3 of the *Cdic* gene became juxtaposed with the 3' untranslated region of the *AnnX* gene, which may be represented as Telomere · · · –*AnnX*–[*Cdic*–*AnnX*]–*Cdic*– · · · Centromere (where again transcription takes place from right to left and the square brackets represent the gene fusion). This [*Cdic*–*AnnX*] fusion was the nascent novel *Sdic* gene, which after additional evolutionary refinement, became tandemly duplicated approximately tenfold,[11] yielding its present situation in the genome as Telomere · · · –*AnnX*–[*Sdic*]$^{\sim 10}$–*Cdic*– · · · Centromere.[37]

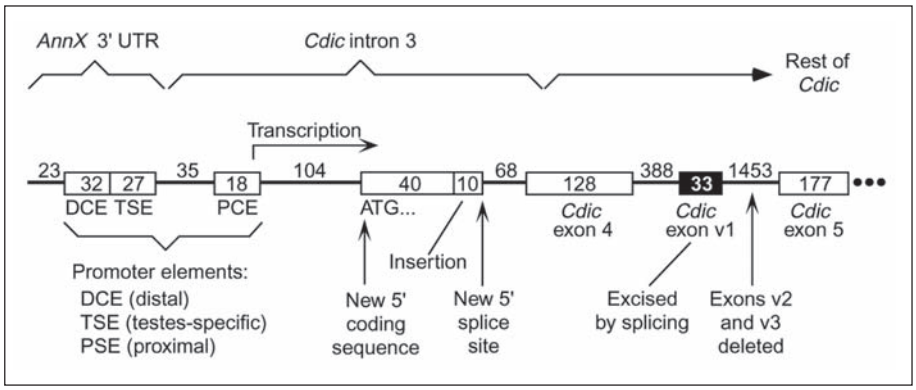

Figure 1. Molecular structure of a portion of one of the analyzed *Sdic* repeats showing the three key promoter elements created by the fusion of the 3' UTR of *AnnX* and intron 3 of *Cdic*. Part of the novel amino end of the Sdic protein derives from *Cdic* intron 3 sequences. Sequence length is indicated in base pairs.

The Molecular Structure of *Sdic*

The reconstituted portion of the *Sdic* repeating unit (in terms of novel promoter and 5' coding regions) is illustrated in Figure 1, in which the gene is oriented so that transcription takes place from left to right. This means that the centromere of the chromosome is far to the left and the telomere of the chromosome is much farther to the right. In each region of the gene, the numbers of nucleotides are indicated. This appears to be the structure of the *Sdic* gene nearest the 5' end of the cluster (nearest to *Cdic*) but there is some variation in sequence and structure from one repeating unit to the next.[45]

The promoter region of *Sdic* is formed from a fusion between the exon for the 3' untranslated region of *AnnX* and intron 3 of *Cdic*. The new promoter shares two similar domains, the distal conserved element (DCE) and the proximal conserved element (PCE), as defined within the wildtype promoter of *Cdic*.[36] The similarity appears to be fortuitous, since neither the *Sdic* DCE nor the *Sdic* PCE are derived from the *Cdic* promoter. Indeed, the *Sdic* DCE derived from the *AnnX* 3' UTR matches the *Cdic* promoter DCE in 25 out of 34 base pairs (bp). The *Sdic* PCE is derived from *Cdic* intron 3 but matches the *Cdic* promoter PCE in 16/20 bp. Another important component of the *Sdic* promoter is the testis-specific element or TSE. This sequence matches the TSE of the testis-specific *betaTub85D* promoter in 21/27 bp. Yet the *Sdic* TSE appears to derive from the 3' UTR of *AnnX*, in which there is a sequence that matches in 22/27 bp. The *Sdic* promoter is sufficient to drive the testis-specific transcription of a construct encoding the Sdic protein fused to a green fluorescent protein reporter.[37]

Although the Sdic protein includes the carboxyl end of Cdic, it is missing 84 amino acids from the amino-terminal end of Cdic. Instead, the Sdic amino-terminus consists of a novel exon that derives largely from *Cdic* intron 3. The Sdic amino end includes domains that are similar to those at the amino end of axonemal dyneins.[37]

As diagrammed in Figure 1, transcription of *Sdic* begins in the PCE. Translation begins 104 nucleotides downstream with an initiation codon that encodes the novel amino end of the Sdic protein. An insertion of 10 base pairs creates a novel splice site, which serves as a donor site for splicing with the wildtype 3' splice acceptor of *Cdic* exon 4. The variable exons (v1–v3) present in *Cdic* between exons 4 and 5[36] are not present in *Sdic* mRNA; exon v1 is removed by RNA splicing, and exons v2 and v3 have been deleted from the *Sdic* genomic sequence. The alternatively spliced exon 5 (which includes exon v4) is spliced in *Sdic* in the longer mode, as found in *Cdic*. The structure and splicing patterns of *Cdic* and *Sdic* are similar for exons 5, 6, and 7, although there are some additional differences near the carboxyl end of the protein.

Reduced Polymorphism in the Region of *Sdic*

The current molecular structure of *Sdic* suggests that in the course of the evolution of this multigene family there was an initial duplication of the region including *AnnX* and *Cdic*, at least three deletions resulting in the *AnnX–Cdic* gene fusion, two more insertions or deletions including one that created a novel splice junction, 11 nucleotide substitutions including reversal of a chain-terminating codon, and an estimated tenfold tandem reiteration of the newly fashioned *Sdic* gene.[36] All of these mutations and gene fixations have occurred in a relatively short time after the divergence of *D. melanogaster* and *D. simulans*, and evolutionary refinement may still be taking place.

Recent adaptive evolution of *Sdic* might be detectable as a selective sweep, which in principle could be detected as a reduction in the level of genetic polymorphisms in the *Sdic* region and a frequency distribution of genetic variation skewed toward rare alleles. A reduced level of polymorphism in the *Sdic* region was noted in the original report.[37] In particular, the nucleotide sequences of 1200 bp of *Sdic* and 985 bp of *Cdic* from each of nine strains of geographically diverse origin yielded estimates of nucleotide polymorphism (θ) of 1.23E–3 ± 0.83E–3 and 0.78E–3 ± 0.66E–3, respectively, and estimates of nucleotide diversity (π) of 0.89E–3 ± 0.73E–3 and 0.45E–3 ± 0.50E–3, respectively. These are among the lowest estimates of nucleotide variation found in nuclear genes of diverse geographic isolates of *Drosophila*[35] and are consistent with a relatively recent selective sweep in the *Sdic* region.

The Issue of Background Selection

Charlesworth and Charlesworth[14] were quick to point out, correctly, that while a showing of reduced polymorphism is necessary to infer a selective sweep, it is not sufficient. They argued that a reduced level of polymorphism in a region of low recombination, such as at the base of the X-chromosome, is also consistent with background selection due to deleterious mutations. Background selection results from the fact that each new deleterious mutation that occurs dooms some genetically linked region of chromosome to eventual extinction. The lower the rate of recombination, the larger the region of chromosome that is affected. The population effect of any new deleterious mutation is thus to reduce by one the number of chromosomes that the affected region of the genome can contribute to remote future generations. If there is absolute linkage, then the whole chromosome is affected; if there is recombination, then a smaller region flanking the mutation is affected. In either case, a sufficient density of harmful mutations will reduce the number of surviving lineages to such an extent that the degree of polymorphism will be smaller than expected, given the actual population size, and the tighter the linkage the greater the disparity.

Nurminsky and colleagues[38] rejoinder was based on the amount of codon usage bias in the region. In *Drosophila*, highly expressed genes tend to have a biased pattern of codon usage,[49] which apparently results from weak selection that favors more rapid or more accurate translation.[3] Background selection in a region of relatively tight linkage would, owing to the reduction in effective population size, be expected to result in a diminution in codon usage bias in genes across the region. Although the data available at the time showed an extremely sharp increase in codon usage bias as the gene locations proceeded outward from the centromeric heterochromatin of the X-chromosome, the complete genomic sequence of *D. melanogaster*[1] reveals a less dramatic pattern. Figure 2 shows the codon usage bias of 201 genes at the base of the X-chromosome, oriented with the centromere off to the right, taken from data compiled by Hey and Kliman.[24] Codon usage bias is scaled according to the effective number of codons, ENC,[58] a scale in which a smaller effective number of codons corresponds to a greater bias in codon usage. There is gradual, statistically significant ($P < 0.01$) decrease in codon usage bias as the gene positions become closer to the centromeric heterochromatin (i.e., towards cytological band 20). This pattern is consistent with an increase in background selection closer to the

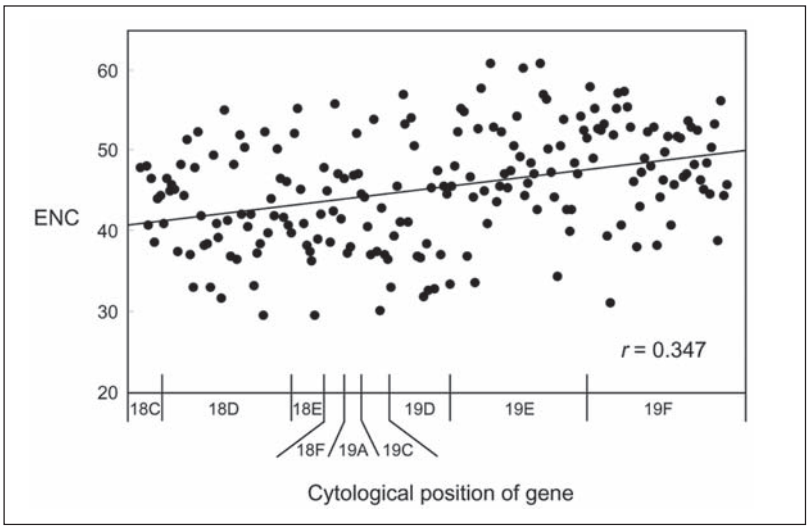

Figure 2. Codon usage bias of genes in the base of the euchromatin of the X-chromosome, oriented with the centromeric heterochromatin off toward the right. The measure of codon bias is the effective number of codons,[58] which scales inversely with codon usage bias. Hence larger values of the ENC are associated with less biased codon usage. Based on data from Hey and Kliman.[24]

centromeric heterochromatin. However, the level of codon usage bias in the *Sdic* region (19A) is not markedly different from that of the *Zw* region (18D). These observations suggest that background selection does have some effect in the *Sdic* region, but not likely a sufficiently strong effect to reduce the level of polymorphism to that observed for *Sdic* and *Cdic*.

Further Evidence for a Selective Sweep

But of course, a general argument based on codon usage bias is indirect and uncertain. A more rigorous analysis was carried out by Nurminsky et al[39] who studied the level of polymorphism of ten genes at the base of the X-chromosome in a worldwide sample of 15 isofemale lines of *D. melanogaster* and 7 isofemale lines of *D. simulans*. The data from *D. simulans* served for comparison and showed a linear decrease in the level of polymorphism as a function of a gene's proximity to the centromeric heterochromatin. The data from *D. melanogaster* revealed a similar trend, but included a statistically significant "dip" in the level of polymorphism in the *Sdic* region. This pattern is entirely consistent with a selective sweep at or close to the *Sdic* locus.

A recent selective sweep was also implied by the frequency spectrum of polymorphisms.[39] In *D. melanogaster*, the frequency spectrum across the base of the X-chromosome was skewed toward rare variants, considering either synonymous polymorphisms only (Wilcoxon signed-rank test $P = 0.04$) or for synonymous and nonsynonymous polymorphisms combined (Wilcoxon signed-rank test $P = 0.01$). The corresponding P-values for the data from *D. simulans* were 0.44 and 0.28, respectively.

More evidence of a selective sweep can be gathered by comparing the *Sdic* locus to its progenitor sequence, *Cdic*. Between these two genes' aligned coding regions, Nurminsky et al[37] found six replacement changes but only two synonymous changes. This higher than average nonsynonymous to synonymous ratio of substitutions suggests that positive Darwinian selection has played a role in the evolution of *Sdic* although a decrease in selective constraints, particularly after a gene duplication event,[40] can also explain this pattern. Further, a surprisingly

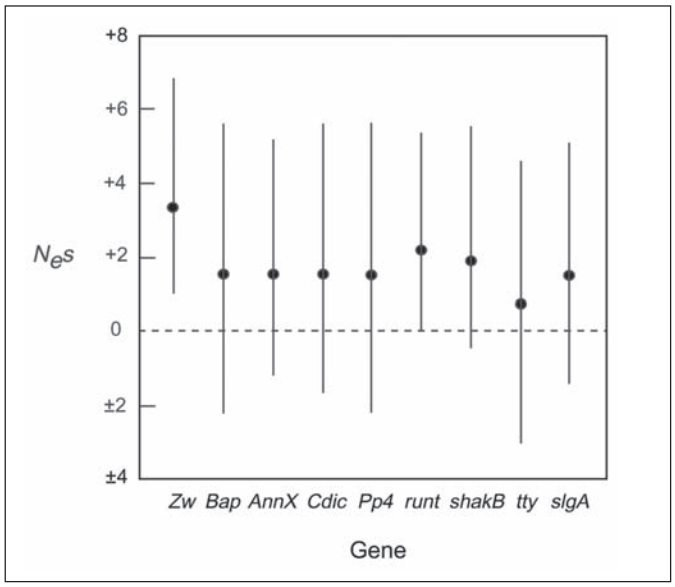

Figure 3. Estimates of the scaled average selection coefficient ($N_e s$) of amino acid replacements, and the 95% credible intervals, for a sample of genes across the base of the X-chromosome in *D. melanogaster* and *D. simulans*, based on the hierarchical Bayesian analysis outlined in Bustamante et al.[13]

complex pattern of deletions in the 3' exon has been recently found among *Sdic* copies and in relation to *Cdic*.[45]

Bayesian Analysis of Polymorphism and Divergence in the *Sdic* Region

Results of a hierarchical Bayesian analysis of polymorphism and divergence of genes across the *Sdic* region is shown in Figure 3, where an estimate of $N_e s$ for each gene and its 95% credible interval is indicated.[13] *Sdic* is not included, since the gene cluster does not exist in *D. simulans*.

To relate the data in Figure 3 to the full analysis of 43 genes in Bustamante et al[13] note that the value of $N_e s$ for *Zw* ranks second highest among the full set of 43 genes, and the values of $N_e s$ for eight of the nine genes in Figure 3 rank in the top 60% of the genes in the full set. Hence, although only two of the genes in Figure 3 (*Zw* and *runt*) have significant values of $N_e s$ by the criterion that their 95% credible intervals do not overlap zero, the generally large values of $N_e s$, averaging 1.73, seem to reflect the apparent action of positive selection across the region. What is not so clear is the extent to which the apparent level of selection indicates a selection at each locus individually as opposed to the effects of genetic linkage with one or two strongly selected genes in the region. Nevertheless, the analysis of polymorphism and divergence reinforces the conclusion reached from the frequency spectrum of synonymous polymorphisms that there has been at least one positively selected sweep in this region. The genetic linkage across the region complicates the interpretation, because the Bayesian analysis assumes that the genes are independent, but on the other hand, any reduction in N_e in the region that results from background selection implies that the values of s are actually greater than the estimated values of $N_e s$ would imply. In any case, the results in Figure 3 suggest to us that there may well have been more than one selective sweep in the region, perhaps in more than one gene, since a selective sweep can impel to fixation only those amino acid replacements with which the favorable mutation happens to be linked.

One interesting sidelight of the data has to do with the effective population size of *D. simulans* relative to *D. melanogaster*. Analysis of synonymous substitutions suggests that N_e for *D. simulans* is larger than that for *D. melanogaster*.[4] Maximum likelihood estimates of the ratio of the effective population sizes in the *Sdic* region yield an estimated ratio of 1.486 (95% confidence interval 0.723–2.249) for all *D. melanogaster* populations taken together. However, when the analysis is restricted to *D. melanogaster* lines from Zimbabwe, the estimated ratio of effective sizes is 0.994 (95% confidence interval 0.581–1.407). These are obviously not significantly different, but they do serve to support the inference that worldwide *D. simulans* has an effective population size about 50% greater than that of *D. melanogaster* and additionally, that there is more genetic variation in African, particularly Zimbabwe, populations of *D. melanogaster* than there is in North American populations.[8]

The higher effective population size found among Zimbabwe lines compared to other global *D. melanogaster* lines, as suggested by the Bayesian analysis, supports this population's distinct, isolated, and presumably stable nature.[5,25,60] More importantly, it presents us with another opportunity to test the selective sweep hypothesis in the *Sdic* region. Once a selective sweep occurs, it takes approximately N_e generations (depending on the strength of selection) for the population to return back to equilibrium.[44] Since the Zimbabwe population has a higher effective population size relative to other more recently diverged *D. melanogaster* populations, deviations from neutrality would be easier to detect. Table 1 shows that although values of the Tajima's *D* statistic are not significantly different from zero, all ten loci (located in the *Sdic* region) with samples solely from Zimbabwe populations of *D. melanogaster*, produce negative Tajima's *D* values. This observed skew in frequency towards rare variants was not found in *D. simulans* nor with other *D. melanogaster* populations and, together with the previously reported pattern of low polymorphism, suggests that a recent sweep(s) has taken place in African *D. melanogaster* populations in or around the *Sdic* locus.

Rapid Evolution of Male-Specific Genes

The accumulated set of observations which include the rapid formation of the *Sdic* gene cluster, the low level of *Sdic* nucleotide diversity and the frequency distribution of rare *Sdic* variants, as well as the observed patterns of variation in genes neighboring the *Sdic* locus—the suppressed levels of genetic variation, the lower than expected decrease in codon bias, the consistently negative Tajima's *D* values in African populations, and the slightly positive selection intensities estimated from the data—together provide strong evidence that a selective sweep, or a series of recurrent sweeps, has taken place at the *Sdic* locus. This inference also fits into the wider context of the faster evolution of male-specific traits, particularly those involved in fertility.[15,61] As a protein expressed specifically in the sperm tail, *Sdic* may be positively selected under a variety of sexual selection mechanisms. For example, sperm competition,[17,18] sexual conflict[46] and sexual coevolution[52] have been demonstrated in *Drosophila* and may be a potent force in the molecular evolution of sperm-specific genes.

Recently, a number of male-specific genes have been identified that, like *Sdic*, possess a high ratio of replacement to silent fixed substitutions.[50] This pattern of high amino acid divergence in male-specific proteins appears to be a general phenomenon among a wide variety of taxa but is especially evident in *Drosophila*.[16,50] For example, many of the most rapidly evolving genes, as revealed by two-dimensional electrophoresis of *Drosophila* proteins, are male-specific.[15,19,53] Other rapidly evolving male-specific genes or genetic systems in *Drosophila* include segregation distortion,[32,59] sex ratio in *D. simulans*,[6] *Mst40*[47] and *Stellate*.[12,34,41]

The rapid evolution of the *Sdic* gene cluster also represents a remarkable example of gene evolution *in statu nascendi*. Interestingly, of the few known examples of incipient gene/domain formation among closely related species, many appear to be associated with male reproductive traits, particularly spermatogenesis. For example, the *jingwei* gene in the *D. teissieri* /*D. yakuba*

Table 1. Tajima's* D *on loci near the Centromeric region of the X-chromosome

			D. melanogaster			*D. simulans*
Locus	**Band**	**Length (bp)**	**Non-Zimbabwe**	**Zimbabwe**	**All Populations**	**Global Populations**
Zw	18D13	537	0.38 (6)	-0.22 (6)	-0.33 (12)	-0.33 (7)
Bap	18E4-18E5	471	0.18 (8)	-0.53 (5)	-0.75 (13)	-1.21 (7)
AnnX	19C1	606	-1.29 (9)	-0.93 (6)	-0.83 (15)	0.85 (7)
Sdic	19C1	1146	-0.58 (9)	-0.47 (5)	0.43 (14)	N.A.
Cdic	19C1	777	-0.84 (9)	-0.19 (6)	0.08 (15)	0.16 (7)
Pp4	19C	339	0.33 (8)	-0.93 (6)	-0.34 (14)	0.69 (7)
run	19E2	585	0.00 (7)	-0.79 (6)	-0.72 (13)	-0.73 (7)
shakB	19E3	705	-1.24 (7)	-0.93 (6)	-1.65 (13)	-1.01 (7)
tty	19F4	594	-0.36 (9)	-0.79 (6)	-0.70 (15)	-0.35 (7)
slgA	19F6-20A1	495	-0.58 (9)	-0.83 (6)	-0.33 (15)	0.71 (6)
su(f)	20E	966	N.A.	N.A.	N.A.	0.21 (7)
		Variance	0.38	0.08	0.32	0.54
		Sign Test	P<0.75	P<0.01**	P<0.11	P=1

N.A.= Data not available. Number of sequences in parentheses. Tajima's *D* is not significantly different from zero in all cases.

lineage[56] has recently evolved and is expressed specifically in the testis. Similarly, *Odysseus*—although not a newly evolved gene—contains rapidly evolving homeodomains involved in sperm function that have been recently fixed solely in *D. mauritiana*, a sibling species in the *D. melanogaster* complex.[54,55] Hence, it appears that while other genetic systems may possess a higher level of selective constraints, spermatogenesis may be more prone to allow for the coopting of novel genes and function. Consequently, the greater potential for selective sweeps may be an intrinsic property of genes expressed in the male reproductive system. Therefore, the observed presence of selective sweep(s) in the *Sdic* region may be the result of the combination of *Sdic*'s location in a tightly linked region of the genome together with its potential fitness consequences on male fertility.

Acknowledgements

This work was supported by NIH grants GM60035 (DH) and GM61549 (DN), NSF grant DMS-0107420 (SAS) and by fellowships from the Natural Sciences and Engineering Council of Canada (RJK), the Marshall-Sherfield fund (CDB), the National Research Council of Spain (JMR), the Foundation for Science and Technology of Portugal (ARP).

References

1. Adams MD, Celniker SE, Holt RA et al. The genome sequence of Drosophila melanogaster. Science 2000; 287:2185-95.
2. Akashi H. Synonymous codon usage in Drosophila melanogaster: Natural selection and translational accuracy. Genetics 1993; 136:927-35.
3. Akashi H. Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA. Genetics 1995; 139:1067-76.

4. Akashi H. Molecular evolution between Drosophila melanogaster and D. simulans: Reduced codon bias, faster rates of amino acid substitution, and larger proteins in D. melanogaster. Genetics 1996; 144:1297-307.
5. Andolfatto P, Przeworski M. Regions of lower crossing over harbor more rare variants in African populations of Drosophila melanogaster. Genetics 2001; 158:657-65.
6. Atlan A, Mercot H, Landre C et al. The sex-ratio trait in Drosophila simulans: Geographical distribution of distortion and resistance. Evolution 1997; 51:1886-95.
7. Barton GJ, Newman RH, Freemont PS et al. Amino acid sequence analysis of the annexin super-gene family of proteins. Eur J Biochem 1991; 198:749-60.
8. Begun DJ, Aquadro CF. African and North American populations of Drosophila melanogaster are very different. Nature 1993; 365:548-50.
9. Begun DJ, Aquadro CF. Evolutionary inferences from DNA variation at the 6-phosphogluconate dehydrogenase locus in natural populations of Drosophila: Selection and geographic differentiation. Genetics 1994; 136:155-71.
10. Benassi V, Depaulis F, Meghlaoui GK et al. Partial sweeping of variation at the Fbp2 locus in a West African population of Drosophila melanogaster. Mol Biol Evol 1999; 16:347-53.
11. Benevolenskaya E, Nurminsky D, Gvozdev V. Structure of the Drosophila melanogaster annexin X gene. DNA Cell Biol 1998; 14:349-57.
12. Bozzetti MP, Massari S, Finelli P et al. The Ste locus, a component of the parasitic cry-ste system of Drosophila melanogaster, encodes a protein that forms crystals in primary spermatocytes and mimics properties of the beta subunit of casein kinase. Proc Natl Acad Sci USA 1995; 92:6067-71.
13. Bustamante CR, Nielsen R, Sawyer SA et al. The cost of inbreeding in Arabidopsis. Nature 2002; 46:531-4.
14. Charlesworth B, Charlesworth D. How was the Sdic gene fixed? Nature 1999; 400:519-20.
15. Civetta A, Singh RS. High divergence of reproductive tract proteins and their association with postzygotic reproductive isolation in Drosophila melanogaster and Drosophila virilis group species. J Mol Evol 1995; 41:1085-95.
16. Civetta A, Singh RS. Broad-sense sexual selection, sex gene pool evolution, and speciation. Genome 1999; 42:1033-41.
17. Civetta A, Clark A. Correlated effects of sperm competition and postmating female mortality. Proc Natl Acad Sci USA 2000; 97:13162-5.
18. Clark AG, Aguade M, Prout T et al. Variation in sperm displacement and its association with accessory gland protein loci in Drosophila melanogaster. Genetics 1995; 139:189-201.
19. Coulthart MB, Singh RS. High level of divergence of male-reproductive-tract proteins between Drosophila melanogaster and its sibling species, D simulans Mol Biol Evol 1988; 5:182-91.
20. Depaulis F, Brazier L, Veuille M. Selective sweep at the Drosophila melanogaster Suppressor of Hairless locus and its association with the In(2L)t inversion polymorphism. Genetics 1999; 152:1017-24.
21. Fay JC, Wyckoff GJ, Wu CI. Testing the neutral theory of molecular evolution with genomic data from Drosophila. Nature 2002; 415:1024-6.
22. Geisow MJ. Annexins: Forms without function but not without fun. Trends Biotechnol 1991; 9:180-1.
23. Hamblin MT, Veuille M. Population structure among African and derived populations of Drosophila simulans: Evidence for ancient subdivision and recent admixture. Genetics 1999; 153:305-17.
24. Hey J, Kliman RM. Interactions between natural selection, recombination and gene density in the genes of Drosophila. Genetics 2002; 160:595-608.
25. Hollocher H, Ting C-T, Wu M-L et al. Incipient speciation by sexual isolation in Drosophila melanogaster: Extensive genetic divergence without reinforcement. Genetics 1997; 147:1191-201.
26. Hudson RR, Bailey K, Skarecky D et al. Evidence for positive selection in the superoxide-dismutase (Sod) region of Drosophila melanogaster. Genetics 1994; 136:1329-40.
27. King SM, Barbarese E, Dillman JF et al. Brain cytoplasmic and flagellar outer arm dyneins share a highly conserved Mr 8,000 light chain. J Biol Chem 1996; 271:19358-66.
28. Kirby DA, Stephan W. Haplotype test reveals departure from neutrality in a segment of the white gene of Drosophila melanogaster. Genetics 1995; 141:1483-90.

29. Kirby DA, Stephan W. Multi-locus selection and the structure of variation at the white gene of Drosophila melanogaster. Genetics 1996; 144:635-45.
30. Labate JA, Biermann CH, Eanes WF. Nucleotide variation at the runt locus in Drosophila melanogaster and Drosophila simulans. Mol Biol Evol 1999; 16:724-31.
31. Maynard Smith J, Haigh J. The hitch-hiking effect of a favorable gene. Genet Res 1974; 23:23-5.
32. McClean JR, Merrill CJ, Powers PA et al. Functional identification of the segregation distorter locus of Drosophila melanogaster by germline transformation. Genetics 1994; 137:201-9.
33. McDonald JH, Kreitman M. Adaptive protein evolution at the Adh locus in Drosophila. Nature 1991; 351:652-4.
34. Mckee BD, Satter MT. Structure of the Y chromosomal Su(Ste) locus in Drosophila melanogaster and evidence for localized recombination among repeats. Genetics 1996; 142:149-61.
35. Moriyama EN, Powell JR. Intraspecific nuclear DNA variation in Drosophila. Mol Biol Evol 1996; 13:261-77.
36. Nurminsky DI, Benevolenskaya EV, Nurminskaya MV et al. Cytoplasmic dynein intermediate chain isoforms with different targeting properties created by tissue-specific alternative splicing. Mol Cell Biol 1998a; 18:6816-25.
37. Nurminsky DI, Nurminskaya MV, De Aguiar D et al. Selective sweep of a newly evolved sperm–specific gene in Drosophila. Nature 1998b; 396:572-5.
38. Nurminsky DI, Hartl DL. How was the Sdic gene fixed? Nature 1999; 400:520.
39. Nurminsky DI, Aguiar DD, Bustamante CD et al. Chromosomal effects of rapid gene evolution in Drosophila melanogaster. Science 2001; 291:128-30.
40. Ohno S, ed. Evolution by Gene Duplication. Berlin: Springer-Verlag, 1970.
41. Palumbo G, Bonaccorsi S, Robbins LG et al. Genetic analysis of stellate elements of Drosophila melanogaster. Genetics 1994; 138:1181-97.
42. Parsch J, Meiklejohn CD, Hartl DL. Patterns of DNA sequence variation suggest the recent action of positive selection in the janus-ocnus region of Drosophila simulans. Genetics 2001; 159:647-57.
43. Paschal BM, Mikami A, Pfister KK et al. Homology of the 74-kD cytoplasmic dynein subunit with a flagellar dynein polypeptide suggests an intracellular targeting function. J Cell Biol 1992; 118:1133-43.
44. Perlitz M, Stephan W. The mean and variance of the number of segregating sites since the last hitchhiking event. J Math Biol 1997; 36:1-23.
45. Ranz JM, Ponce AR, Hartl DL et al. Origin and evolution of a new gene expressed in the Drosophila sperm axoneme. Genetica 2003 188:233-44.
46. Rice WR. Sexually antagonistic male adaptation triggered by experimental arrest of female evolution. Nature 1996; 381:232-4.
47. Russell SRH, Kaiser K. A Drosophila melanogaster chromosome-2L repeat is expressed in the male germ line. Chromosoma 1994; 103:63-72.
48. Sawyer SA, Hartl DL. Population genetics of polymorphism and divergence. Genetics 1992; 132:1161-76.
49. Shields DC, Sharp PM, Higgins DG et al. "Silent" sites in Drosophila genes are not neutral: Evidence of selection among synonymous codons. Mol Biol Evol 1988; 5:704-16.
50. Singh RS, Kulathinal RJ. Sex gene pool evolution and speciation: A new paradigm. Genes Genet Syst 2000; 75:119-30.
51. Smith NGC, EyreWalker A. Adaptive protein evolution in Drosophila. Nature 2002; 415:1022-4.
52. Swanson W, Vacquier V. The rapid evolution of reproductive proteins. Nature Reviews Genetics 2002; 3:137-44.
53. Thomas S, Singh RS. A comprehensive study of genetic variation in natural population of Drosophila melanogaster. VII. Varying rates of genic divergence as revealed by two-dimensional electrophoresis. Mol Biol Evol 1992; 9:507-25.
54. Ting C-T, Tsaur S-C, Wu M-L et al. A rapidly evolving homeobox at the site of a hybrid sterility gene. Science 1998; 282:1501-4.
55. Ting C-T, Tsaur S-C, Wu C-I. The phylogeny of closely related species as revealed by the genealogy of a speciation gene, Odysseus. Proc Natl Acad Sci USA 2000; 97:5313-6.

56. Wang W, Zhang JM, Alvarez C et al. The origin of the Jingwei gene and the complex modular structure of its parental gene, yellow emperor, in Drosophila melanogaster. Mol Biol Evol 2000; 17:1294-301.
57. Wiehe THE, Stephan S. Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from Drosophila melanogaster. Mol Biol Evol 1993; 10:842-54.
58. Wright F. The 'effective number of codons' used in a gene. Gene 1990; 87:23-9.
59. Wu C-I, Lyttle TW, Wu M-L et al. Association between a satellite DNA sequence and the Responder of Segregation Distorter in D. melanogaster. Cell 1988; 54:179-89.
60. Wu C-I, Hollocher H, Begun D et al. Sexual isolation in Drosophila melanogaster: A possible case of incipient speciation. Proc Natl Acad Sci USA 1995; 92:2519-23.
61. Wu C-I, Davis AW. Evolution of postmating reproductive isolation: The composite nature of Haldane's rule and its genetic bases. Am Nat 1993; 142:187-212.

Chapter 4

Detecting Selective Sweeps with Haplotype Tests:
Hitchhiking and Haplotype Tests

Frantz Depaulis, Sylvain Mousset and Michel Veuille

Abstract

In this chapter, neutrality tests based on haplotype distribution are evaluated as a way of detecting selective sweeps. Several kinds of haplotype tests are reviewed, including haplotype number, haplotype diversity and haplotype partition tests. We focus on incomplete sweeps, where recombination between the selected locus and a given marker allows for several preexisting neutral lineages to survive the sweep and for some preexisting genetic variation to remain in a sample. Several problems are addressed, including the distinction between possible alternative hypotheses, the effect of sampling strategy, of conditioning the statistics on the population mutational parameter θ and/or the observed number of polymorphic sites S and, finally, the effect of intragenic recombination together with the choice of one- vs. two-tailed tests. Corresponding guidelines are proposed. To compare the power of haplotype tests and of other classical tests to detect selective sweeps, we use a simple selective sweep model with a deterministic approximation, allowing for genetic exchange between the selected locus and a given neutral marker. We conclude that there are ways of overcoming the difficulties in applying the tests, which are powerful means for revealing incomplete selective sweep effects.

Introduction

Since the proposal of the neutral theory,[1] scientists have been looking for the footprint of phenotypic selection at the molecular level and have found frequent departures from the simple neutral model.[2] An indirect way of detecting potentially rare selective events is to use their effect on neighboring variation, the "hitchhiking" effect (see ref. 3 for a theoretical review). In the usual restricted sense, this refers to the effect of advantageous mutations on linked neutral variation, which is also called the "selective sweep". After the fixation of an advantageous mutation, the most obvious hitchhiking effect is a reduction of genetic variation at linked neutral loci. A potential proof of these effects was the discovery of a genome-wide correlation between levels of polymorphism and local recombination rate in *Drosophila melanogaster*.[4-6] This trend was subsequently confirmed in humans,[7] mice[8] and tomatoes.[9] Since such a correlation is not found at the divergence level, it is in disconsent with neutral mutational expectations (but see ref. 10). Hitchhiking has been proposed as a possible process behind this observation. Assuming a uniform input of advantageous mutations along chromosomes, selective sweeps are

Selective Sweep, edited by Dmitry Nurminsky. ©2005 Eurekah.com and Kluwer Academic/Plenum Publishers.

expected to extend further and show stronger effects in low recombination regions of the genome. However, the background selection—the effect of the removal of deleterious mutations on linked neutral variation—is an alternative mechanism that can account for this pattern.[11,12]

A potential way of distinguishing between these two models is that selective sweeps predict a shift in the frequency spectrum of mutations towards an excess of low frequency variants when compared to neutral predictions.[13] We will describe these effects in terms of genealogical pattern, with reference to the coalescent theory, which represents the genetic history of a sample by gene genealogies[14] (Fig. 1a). It provides an intuitive way of looking at historical perturbation effects. In the simplest case of a complete selective sweep with no recombination, during a transient period of recovery of polymorphism after the selective perturbation, the genealogy of the marker considered is star-like, leading to an excess of rare variants (Fig. 1c). This is not true

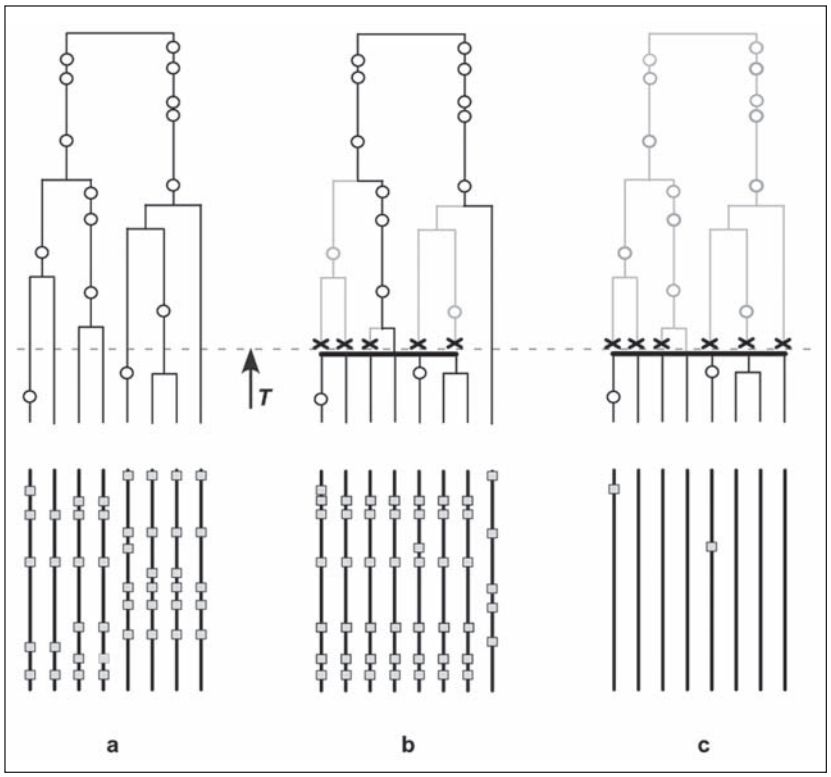

Figure 1. Outline of the shape of genealogies (n=8) relevant to various models. Mutations are indicated on the tree as circles (top panels), and resulting polymorphic sites are schematically shown as squares on the corresponding alleles below. T: age of the selective sweep. a) Neutral genealogy, constant size population (S=15, K=6, H=0.81). b) Genealogy after a hitchhiking with recombination (S=13, K=4, H=0.56). The thick horizontal line represents the lineage carrying the beneficial mutation when it arises. Most of the other lineages become extinct being outcompeted by the advantageous allele, but one survives the sweep through recombination with the advantageous mutation. Haplotype diversity is drastically reduced due to the high frequency of the most common haplotype. Note also the excess of derived mutations in high frequency. c) After a hitchhiking without recombination, a single lineage survives the selective sweep ($\alpha \to \infty$, S=2, K=3, H=0.41). Though all mutations are unique and the number of haplotypes is maximal, there is no power to detect the recent sweep due to the very low polymorphism in the sample.

in the case of background selection, at least for its original form involving strongly deleterious mutations.[11] The result of background selection can be approximated by a reduction of the effective population size, but the remaining variation and the corresponding trees show a neutral distribution[11,12] (Fig. 1). Empirical attempts to distinguish between the two hypotheses initially focused on low recombination regions of the genome, such as the subtelomeric regions and the tiny fourth chromosome of *D. melanogaster*.[4,6,15] However, in these regions variation is reduced to such an extent that there is hardly any information left for testing neutrality, except through extensive genotyping.[16,17]

An alternative approach is to focus on less strong hitchhiking effects which, while partially preserving the preexisting variation, disrupt its pattern in the population. The possible underlying processes include an incomplete sweep of an advantageous mutation on its way to fixation, a recent balanced polymorphism, fluctuating selection,[3] or interference between several adaptive variants on their way to fixation (the "traffic" hypothesis[18]). Weak hitchhiking could also simply reflect the loose linkage between the marker and the selected locus relative to the strength of selection, thus allowing for recombination events between the two loci during the selective stage[19] (hereafter "hitchhiking with recombination"). In genealogical terms, this would prevent several preexisting neutral lineages from being completely removed during the sweep (Fig 1b).

To detect such effects, among others, a number of statistical tests have been proposed. These are often called "neutrality tests" for short, but they really test a full neutral model with all its assumptions. These tests are based on the Wright-Fisher model, which assumes the neutrality of mutations, and also has strong demographic assumptions in a broad sense (panmictic isolated population of constant size, at mutation-drift equilibrium with a Poisson distribution of the offspring). The tests also rely on a mutational model, where the choice of a particular model depends on the type of genetic marker(s) considered. For the nucleotide variation, which is the focus of our review, the infinite site model (ISM)[1] seems to show the best fit. In its original form, the ISM assumes the intragenic recombination absent, the mutation rate unchanged with time and uniform along the sequence, and the mutations mutually independent. In addition, it assumes that each mutation affects a new site, hence there is no possible homoplasy. As a consequence, the ISM is the most powerful model to be tested for, or more generally to make inferences from – provided that there are such informative markers showing variation in the species considered. Nevertheless, when a significant departure from the neutral model is detected, this can be due to violation of any of the model assumptions, including selective neutrality as well as the demographic and mutational effects. Selective and demographic perturbations predict similar effects on a single locus.[3,19] However, when multiple loci are analyzed, they are expected to show similar effect of demographic perturbations, but they would likely differ in respect to the selective events.

The first class of neutrality tests proposed for detecting departure from the neutral model focused on the frequency spectrum of mutations.[20-22] This approach makes no use of the information contained in the association of different polymorphic sites, which may be informative about the underlying tree and about the events that could have shaped it[19] (Fig. 1). Simple summary statistics using this information rely on the distribution of haplotypes. These statistics, which are the basis of several neutrality tests, are the subject of the present chapter. We use the term "haplotype structuring" to describe any pattern characterized by an excess of linkage disequilibrium structure as compared to the standard neutral model, including a deficit in the number of haplotypes or in haplotype diversity, and an excess of high frequency haplotypes. These effects can be found separately or in various combinations. Hitchhiking increases linkage disequilibrium, and thus haplotype structuring, by shifting haplotypes to high frequencies[23] (Fig. 1b). If the last selective sweep perturbation is recent, there would have been little time for intragenic recombination to alter this structuring.

Available Haplotype Tests

Haplotype Partition Test HP

This haplotype test was proposed by Hudson et al.[24] It computes the probability of finding a subset of at least n_p sequences with no more than S_p polymorphic sites given that the total sample of n sequences shows S polymorphic sites (see Table 1 for the definitions of all symbols used and Table 2 for a summary of the characteristics of the tests). Briefly, it tests for the occurrence of a major "haplotype class", i.e., a subset of sequences with low variation as compared to the total sample. This test (hereafter *HP*) was originally proposed and applied on data *a posteriori*: the unusual pattern was revealed first, then the corresponding test was designed; finally, n_p and S_p were chosen arbitrarily according to the data. However, the probabilities of observing the pattern by chance were so low that the qualitative conclusions were unlikely to be affected by tailoring the test to the dataset.

Haplotype Number K

More general alternative tests relying on the distribution of haplotypes were subsequently proposed. The haplotype number statistics were considered independently with slightly different approaches by several authors (Table 2).[25-29] The $\boldsymbol{S}$,[25] W,[26] and F_s[27] tests are one-tailed. The F_s statistics assesses an excess of haplotypes, whereas W and $\boldsymbol{S}$ assess a deficit of haplotypes. The classical view is that the number of haplotypes adds little information to the sequence data, under an infinite site model, as it tends to be close to its maximum possible value - either S+1 or n, whichever is the smallest.[30,31] A number of factors can explain this effect, including both "too high" level of variation as compared to the sample size, and the reverse situation of the "too low" level of variation. (Here, the level of variation is scaled by the population mutational parameter $\theta=4N_e\mu$, where N_e is the effective population size and μ the per locus neutral mutation rate). Such parameter values may be a common trend for datasets derived from the species such as human, where large sample sizes are the norm and levels of variation are usually low. However, between these extremes lies a broad range of parameter values encountered in various organisms including *Drosophila*, where the confidence interval of the haplotype number does not include the maximum possible value (Fig. 2a).

Table 1. Definition of the symbols used for parameters

Definitions	Symbols
Sample size	n
Subsample size	n_p
Number of segregating sites in a sample	S
Number of segregating sites in a subsample	S_p
Window size expressed in number of polymorphic sites	S_w
Neutral mutation rate	μ
Effective population size	N
Population mutational parameter[a]	$\theta=4N\mu$
Age of a selective sweep[a]	T_s
Selection coefficient[a]	$\alpha=4Ns$
Rate of genetic exchange between neutral and selected marker[a]	$C=4Nc$
Instantaneous frequency of the advantageous mutation	x
Initial frequency of the advantageous mutation	ε

[a] scaled in unit of $4N$ generations

Table 2. Summary of the characteristics of the tests used

Neutrality Test	Symbol	Direction[a]	Conditional Parameter		Analytic/ Simulations[d]	Robust to Recombination[e]	References
				(Estimator)			
Tajima's *D*	*D*	Both			β approximation	Yes	20
Fu and Li's *D*	D_{fl}	Both	θ		Simulations	Yes	21
Fay and Wu's *H*	H_{fw}	Both (potentially)	*S*		Simulations	Yes	22
Haplotype number	*W*	Deficit	θ	(θ_W[b])	Analytic	Yes	26
	S	Deficit	θ	(π [c])	Analytic	Yes	25
	F_s	Excess	θ	(π [c])	Analytic	No	27
	K	Both	*S*		Simulations	Lower limit	28
Haplotype diversity	*H*	Both	*S*		Simulations	Lower limit	28
Haplotype partition	*HP*	Both (potentially)	*S*		Simulations	Upper limit (of n_p)	24
Pairwise correlation	Z_{nS}	Both (potentially)	*S*		Simulations	Upper limit	33

[a] One vs. two tailed tests; [b] Watterson's[40] estimator; [c] Tajima's[20] estimator; [d] method originally used to compute *P* values (the analytic method can be used only in models without recombination); [e] when *P* values are computed in a model without recombination despite some recombination occurring on the tested locus.

Haplotype Diversity H

A related test considers haplotype diversity $H = 1 - \Sigma\, p_i^2$ where p_i stands for the relative frequency of haplotype i in the sample.[28] It could be corrected for sample size to give an unbiased estimate of the population haplotype diversity. But, whatever the case, the test is conditional on the sample size. The H test is similar to the homozygosity test[32] except that the latter is conditional on both K and θ and therefore uses only the information on the frequency spectrum of alleles (i.e., of haplotypes) and not that of the number of haplotypes.

The three above tests are not independent: low number of haplotypes tends to be correlated with a large subset that has little variation and with low haplotype diversity, the condition which we sum up as "strong haplotype structure". This has to be qualified in that the H test is highly sensitive to the frequency spectrum of haplotypes and thus, like *HP*, it is mainly sensitive to a high frequency of the major haplotype, as predicted under hitchhiking with recombination (Fig. 1b).[19]

There are tests that consider other aspects of linkage disequilibrium structure such as the Z_{nS} test, based on the average pairwise allelic correlation coefficient.[33] The B and Q tests are

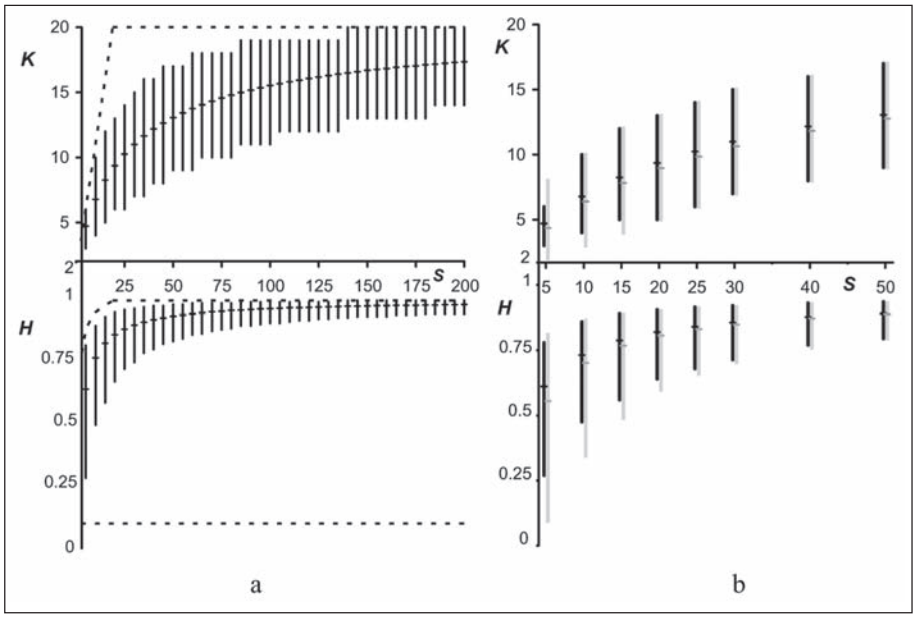

Figure 2. confidence intervals of K (top) and H (bottom) statistics as a function of the number of polymorphic sites S (n=20). Unless otherwise stated, all the simulation results were obtained from 100,000 runs for each set of parameter values following standard coalescent procedures.[14] a) Conditioning of confidence intervals on S. The expectation is indicated by dashes and the maximal and the minimal possible values (in the absence of recombination) by dashed lines (i.e., $K_{min} = 2$ and $K_{max} = min\,(\,n\,, S+1\,)$; $H_{min} = 2\,(\,n-1\,)\,/\,n^2$ and $H_{max} = 1 - (\,n\,(\,1 + 2a\,) - K_{max}\,(\,a + a^2\,)\,)/\,n^2$ where $a = [\,n\,/\,K_{max}\,]$ and $[\,X\,]$ stands for the largest integer below or equal to X; H_{max} simplifies to $1\text{-}1/n$ for $n \geq S + 1$). For a wide range of parameter values, the confidence intervals do not include the maximal and the minimal values, especially for the minimal value. Note that the step-like behavior derives from the discreteness of the statistics rather than from an imprecision due to the simulation approach. b) Confidence intervals conditional on S (black) or θ (grey) as a function of the number of polymorphic sites (or the corresponding value of Watterson's[40] estimate). The simulations with conditioning on θ tend to show broader confidence intervals, especially for low levels of variation.

based on the proportions of congruent adjacent pairs of segregating sites, i.e., pairs showing no evidence of recombination events.[34] Here, we focus on haplotype statistics.

Little is known about their properties. There are several difficulties that we will discuss, and we will propose corresponding guidelines for their use whenever appropriate.

Alternative Hypotheses

Hitchhiking is not the only perturbation that predicts haplotype structuring (Table 3). For a given number of polymorphic sites S, any process that tends to increase the relative length of internal branches as compared to neutral expectations results in excess of haplotype structuring (Fig. 1b). This includes moderate bottlenecks,[35] incomplete sweeps,[24] balanced polymorphisms[36] and population structure.[34] Conversely, any process leading to a star-like tree (a severe bottleneck, a selective sweep without recombination) predicts an excess of haplotypes, a deficit of haplotype structuring (Fig. 1c). Indeed, for a single locus, a bottleneck model predicts results very similar to selective sweep.[19] A severe bottleneck, where all lineages coalesce during the demographic crash, closely matches a selective sweep without recombination. On the other hand, a moderate bottleneck, where several lineages survive the crash without coalescing, predicts results similar to a selective sweep with recombination (unpublished results). Other indications, such as between-loci comparisons, are thus required to distinguish these events (see concluding remarks).[37,38]

Conditioning on S vs. θ

The distribution of these statistics is highly dependent on the levels of genetic variation. Two different approaches have been used for conditioning the tests on a fixed level of variation. The classical approach chosen by Strobeck[25] and Fu[26,27] involves conditioning on the population mutational parameter $\theta = 4N_e\mu$. In this case, the distribution of haplotypes follows the infinite allele model and Ewens's distribution.[39] It could thus be computed exactly by analytical means. In practice, however, in the absence of strong prior knowledge of θ, this population parameter has to be replaced by a point estimate. The S[25] and F_s[27] statistics use Tajima's estimator[20] (the mean pairwise diversity) whereas W[26] uses Watterson's estimator,[40] based on the observed number of segregating sites (often called θ for brevity). This results in nonexact tests, "an achieved level of significance",[26] that can differ substantially from the chosen rejection level and the corresponding tests may not be conservative (e.g., for F_s[27] and S[25] depending on the parameter values). On the contrary, they may equally be too conservative, thereby reducing the power of the tests.[26]

An alternative approach is to condition on S, without any explicit assumption about mutation rate, and to use simulations to obtain the distribution of haplotypes.[24,28,29] The coalescent theory provides a simple framework to obtain the distribution of any statistics that describes polymorphism in a sample empirically, by simulations.[14] This approach is very efficient because only the history of a sample is considered. Hence, not all generations need to be explicitly represented, but only those which have witnessed some events (common ancestry, recombination) that could affect the history of the sample. As it is a sampling theory, it allows for direct comparisons with empirical data and statistical testing.

The rationale of conditioning on S is that θ has to be estimated from the data, whereas S is known with certainty (assuming no homoplasy, a reasonable assumption in general at the intraspecific level). The estimate of θ could be easily skewed by preceding nonneutral events, for example, the level of variation could have been reduced by a selective sweep. Thus, an independent estimate of θ would ideally be needed, but unfortunately this is rarely possible in practice. No datasets respecting all the assumptions of the standard neutral model are generally available to provide such an estimate (which would still show a large degree of uncertainty).

Table 3. Various selective and demographic effects on polymorphism distribution

Perturbation	Strength	Tree Shape	Frequency Spectrum	Haplotype Structure
Background selection	Strong	Neutral like (Fig. 1a)	Weak, excess of rare mutations	Weak, excess of haplotypes
Bottleneck[a] Selective sweep	Strong Complete	Star-like (Fig. 1d)	Excess of rare mutations	Excess of haplotypes, of haplotype diversity, too rare major haplotype
Selective sweep	Partial[b]	Long internal branches, unbalanced (Fig. 1c)	Excess of rare derived mutations	Deficit of haplotypes, of haplotype diversity, too frequent major haplotype class
Bottleneck Population structure	Moderate	Long internal branches (Fig. 1b)	Deficit of rare mutations	Deficit of haplotypes, of haplotype diversity
Balanced selection	Old			

[a] Could also refer to population expansion which shows similar effects; [b] Refers to a sweep in the process, hitchhiking with recombination, fluctuating selection, interference between mutations or recent balanced selection

Moreover, the parameter θ depends on the sampling scheme: which population or set of populations is being sampled. It also depends on the locus considered: the neutral mutation rate depends on the raw total mutation rate, but also on the fraction of deleterious mutations. This fraction needs to be subtracted from the raw rate to get the neutral rate. It thus depends on the level of constraint of the locus and on its composition in terms of exons, introns, of functional domains of the protein. Finally, even if the true value of θ was known with certainty, conditioning on θ leads to a broader confidence interval as compared to S, thereby reducing the power of the tests (Fig. 2b).

Conditioning on both θ and S, employing for example a rejection algorithm or a more efficient importance sampling method, has been proposed.[41] However, this approach implies the use of nonrandom subset of genealogies from the original neutral distribution and thus cannot be strictly regarded as a neutrality test. For instance, consider a sample of size n=20, with an observed number of mutations S=14, with a number of haplotypes K=11 derived from a population with a mutational parameter θ=28.2 (corresponding to an expected S value of 100). The observed S and K values are unexpectedly low given θ, ($P(S \leq 14|\theta=28.2)<10^{-5}$; $P(K \leq 11|\ \theta=28.2)=0.02$). On the other hand, the observed number of haplotypes K is unexpectedly high given the observed S value ($P(K \geq 11|S=14)=0.02$). If we use the rejection algorithm on this set of parameter values, we would have to artificially reject most (> 99.9%) of the neutral genealogies retaining only star-like ones, and to accept the null hypothesis ($P(K \geq 11|S=14, \theta=28.2)=0.15$), despite the obvious inconsistency between these three parameter values. Readers interested in a more thorough discussion of this issue and in validation of the approach conditional on S should read refs. 41-43.

Intragenic Recombination

In this section, we refer to recombination occurring within the sequenced region. This should not be confused with the recombination that may occur between the marker and the selected site, which acts at a larger scale, is implemented differently in the models and has very different consequences. Intragenic recombination generally reduces the variance of statistics, but does not affect the expectation of statistics based, for example, on the frequency spectrum of mutations.[20-22] Corresponding frequency spectrum tests applied without taking into account possible recombination are thus conservative.[34] In contrast, recombination should drastically affect the expectation of haplotype statistics, since it tends to produce additional haplotypes. Indeed, haplotype number is a summary statistic which may be used as an estimator of the population recombination rate.[44] Additional caution should be exercised in applying the haplotype tests when recombination is likely to have occurred in the history of the data (and most recombination events may not be detectable).[45] The lower limit of most tests is, however, conservative towards recombination (except for *Q*).[34] We confirmed this result for the *K* and the *H* tests (results not shown). Tests not taking into account recombination can thus be applied as one-tailed tests if the alternative hypothesis predicts an increased haplotype structure (a deficit of haplotype number, of haplotype heterozygosity). On the other hand, tests that focus on the upper range of haplotype number, such as the F_s test,[27] are not conservative with respect to the occurrence of recombination.

A side effect of using conservative tests is a reduction in power[34] (but see the power section below). Versions of the tests involving recombination have therefore been proposed.[24,29,46,47] In this case, the distribution of haplotypes can only be obtained by simulations, regardless of whether the tests are conditional on *S* or θ. This requires using a conservative value for the recombination rate. Depending on the alternative hypothesis, this has to be a lower bound or an upper bound. In the specific case of hitchhiking (or bottlenecks), this raises the difficulty that, depending on the strength of the effect, this approach predicts either a deficit or an excess of haplotypes (Fig. 1b vs. 1c; see below). In principle, it is advisable to use an estimate of recombination rate derived independently of the data (e.g., from classical genetics), rather than an indirect estimate derived from the data using population genetic tools that assume neutral equilibrium. However, there are substantial genome-wide discrepancies between these two kinds of estimates.[48-50] A very conservative approach is to use the recombination rate that leads to at least as many recombination events as can be inferred from the data using the four gamete rule in at most 5% of the simulations.[29,34,45]

Sampling Strategy and Sliding Window

A difficulty with the haplotype structure tests is that they are sensitive to the sampling strategy: using long sequences compared to the sample size or, conversely, a large sample size compared to the length of the sequence (in terms of θ, *Nr* values) reduces the power of the tests (Fig. 2a). In some cases, the haplotype structure is restricted to a limited region of the sequence and the tests cannot be applied *a posteriori* to this particular region alone. To overcome this difficulty, sliding window versions of the tests have been proposed and applied.[29,46,51] For instance, in the approach conditional on *S*, a window of a given size S_w (expressed in number of polymorphic sites) slides along the sequence from polymorphic site to polymorphic site. The test is applied to each window position *k* and the corresponding "observed" $P_{k\text{-}obs}$ values are stored. Because of multiple testing, these *P* values need to be corrected. They are obviously not independent and a Bonferroni correction would be far too conservative. In a second step, a new set of simulations with the whole dataset parameter values (*n, S*) is run and the test is applied to these simulated datasets with the sliding window approach, exactly as for the actual sequence data. The corresponding minimum *P* values obtained by sliding the window along each

simulated dataset ($P_{min\text{-}sim}$) are used to empirically construct a distribution of the minimum P values that are expected while sliding along the sequence. This distribution is then compared to the observed P values ($P_{k\text{-}obs}$) for sequence data, to derive corrected P values. Refinements of this procedure include correcting for different window sizes[29] in a procedure similar to the S_{max} of ref. 23, focusing on the specific case of two haplotypes. Another refinement allows for different levels of subsample polymorphism (S_p parameter) for *HP*.[46] Note that some tests involving the sliding window approach without taking into account recombination appear to be slightly nonconservative with regard to recombination.[29] Since the sliding window approach is similar to probing a set of different loci, it also allows us to distinguish selective perturbations, which may affect only a particular subregion, from populational perturbations such as the bottleneck, which is expected to affect all subregions of the locus to a similar extent. In this aspect, this approach is similar to that of ref. 23. However, the sliding window technique can be computationally demanding and its effect on the power of the tests is unknown.

Power

In this section, we will address the capacity of various tests to detect several kinds of perturbation from the neutral model and in particular, selective sweep effects.

Previous Results

Fu[26,27] investigated the power of several tests when faced with various hypotheses in a model without recombination. The evaluated tests included haplotype number tests and tests based on the frequency spectrum of mutations such as Tajima's D.[20]

Alternative Hypotheses Involving an Excess of Haplotypes

Alternative hypotheses involving an excess of rare variants and excess of haplotypes included a population growth model and a hitchhiking model without recombination, assuming a deterministic approximation for the increase in frequency of the advantageous mutation. In our nomenclature, both models would be similar to hitchhiking without recombination (Fig. 1c). The results of the two models were again similar and showed F_s to be by far the most powerful test when faced with these alternative hypotheses. On the other hand, when challenged by the background selection model, F_s tends to be more powerful than Tajima's D,[20] but less powerful than the Fu and Li[21] statistics. The latter distinguishes unique versus nonunique mutations, and the reason for this test to be so effective is that background selection affects primarily nonunique mutations.

Alternative Hypotheses Involving a Deficit of Haplotypes

Fu[26] also examined the power of the tests faced with alternative hypotheses leading to unexpectedly long internal branches relative to the length of external ones, thus resulting in an excess of intermediate frequency mutations and in a deficit in the number of haplotypes. In particular, he considered population structure models, based on the infinite island model with various sampling schemes, and the effect of a reduction in population size. Such models would have a similar effect to that of hitchhiking with recombination (Fig. 1b). The haplotype number statistic W generally performed as well or better than other tests which use the frequency spectrum of mutations, and was substantially more powerful than Tajima's D test, for instance. The $\boldsymbol{S}$ test was also found to be powerful when dealing with intermediate migration rate.[25] When the models involved mutation-selection equilibrium, e.g., the ones with balanced selection or deleterious mutations, haplotype number statistics showed power comparable to other tests, depending on the equilibrium frequency of the selected allele and on the estimator of θ used.[27]

Intragenic Recombination Effect

Wall[34] investigated the power of neutrality tests to detect population structure, with an emphasis on intragenic recombination effects. His study surveyed various tests which deal with frequency spectrum of haplotypes and with linkage disequilibrium structure. In agreement with Fu[26] he found that, in the absence of recombination, the haplotype number test generally shows the highest power, provided that individuals are sampled from the same population (which is advisable when applying a neutrality test). But if intragenic recombination occurs, a test that does not take into account recombination becomes by far too conservative as the recombination rate parameter of the population increases. It is then preferable to use tests taking into account recombination, using a conservative recombination rate. Even if the exact recombination rate is known, there is a substantial loss of power with increasing recombination rate.[34] This may reflect the saturation effect of the length of the sequence on haplotype number. The confidence interval tends to be restricted to values close to the maximal possible value for long sequences in terms of recombination rate and of the number of polymorphic sites (Fig. 2a). It is then advisable to shift the trade-off "length of the sequence vs. sample size" in the direction of a larger number of sequenced individuals with shorter sequences.

Selective Sweeps

There are a wide variety of hitchhiking models (reviewed in ref. 3) that could be used, depending on the form of selection (dominance level, frequency dependence, density dependence, fluctuating selection etc.). The results would probably barely differ for a brief selective stage (i.e., for strong selection coefficients), and it is not clear how selection really works in nature. We used a classic selective sweep model, following a procedure similar to that in ref. 13. We adapted their "equilibrium" procedure (using random variables for the selective parameters) to a "single event" procedure as in ref. 22, using the fixed selective parameter values: we assumed exactly one selective sweep event during the history of the genealogy, and we fixed its age. Choosing between equilibrium and single event models depends on the sampling strategy. If loci are sampled randomly with regard to selection, the approach that may require sequencing intergenic regions, equilibrium models may be more appropriate. But these models include additional assumptions about the homogeneity of selection and involve a largely unknown parameter: the frequency of selective sweeps. On the contrary, if, as often, the sampled marker is a candidate for selection in the marker itself or in the neighboring region, the single sweep model may be more appropriate. Assuming a single selective event of a given age also makes the results easier to interpret, as compared to the case where several events of various ages have affected the history of the sample. The selective sweep model approximates the change in frequency of the advantageous allele x, from the virtually null frequency ε (the advantageous mutant is assumed to appear at this point) through the virtual fixation ($x = 1-\varepsilon$), using deterministic equations. During the selective stage, two allelic classes are present in the population, the neutral and the advantageous. Several kinds of events can be considered. Coalescence can only occur between two genes from the same allelic class, and the rate depends on the allele frequency, i.e., $2/x$ between two advantageous alleles (all time scaled parameters and rates are expressed in units of $4N$ generations). Genetic exchange between the selected site and the marker can also occur with a rate $C=4Nc$ (per $4N$ generations; representing either recombination and/or gene conversion rate). Such events lead to gene flow between the two allelic classes: a marker formerly linked to the neutral site can "move" into the advantageous allelic class with the rate of Cx. This should not be confused with intragenic recombination, which generally acts on a smaller scale and has additional effects on haplotype statistics (i.e., disrupts haplotype structure). The selective stage is partitioned into a large number (1,000) of time step increments to take into account the change in the probabilities of the events with x. Complementary

probabilities of the absence of event are multiplied across time steps until the product is less than a uniform random deviate. This point determines the occurrence of an event, chosen according to its relative probability at the current time step. Before or after the selective sweep, all the genes that belong to the same allelic class (neutral or advantageous respectively) show the same fitness and obey the standard neutral coalescent. Unless otherwise stated, we used a strong selection parameter value $\alpha = 4Ns = 10{,}000$ (corresponding to a selection coefficient about $s=0.0025$ for *Drosophila*) leading to a virtually instantaneous selective stage, still allowing for the gene flow for a correspondingly high rate of genetic exchange *C*. What exactly happens during the selective stage probably depends on the details of the model. We use a generic model with a brief selective stage, just to predict how many lineages survive the sweep through recombination-mediated import into the advantageous allelic class and what is the increase in their frequency after the sweep.

In the following power analysis, for simplicity, we consider a version of the *HP* test that is restricted to the absence of polymorphism within the subsample ($S_p=0$). In this case, *HP* test deals with the frequency of the major haplotype. For comparison with other kinds of approaches, we present results for Tajima's D[20] test, as an example of statistics based on the frequency spectrum of mutations. We also used the related Fu and Li's D[21] as an example of a statistics using polarized mutations (assuming that the state—ancestral or derived—of mutations is known, e.g., through the use of an outgroup). The Fu and Li's statistics analyzes the relative proportion of unique, derived mutations. Finally, we also show the results of Fay and Wu's H[22] which is highly dependent on derived mutations of high frequency and was designed specifically to detect selective sweeps. In fact, the different frequency spectrum statistics seem to use substantially the same source of information, to provide similar results and thus appear largely redundant.[27] As another indicator of general linkage disequilibrium structure, we also show the power of Z_{nS}[33] the average pairwise allelic correlation. Like other linkage disequilibrium statistics, it is sensitive to the frequencies of mutations and, therefore, is partly correlated with frequency spectrum statistics. Since departure from the neutral model can be observed in both directions, depending on the parameter values, we show separately the two directions of departure, but for clarity we present only the curves showing a power above the chosen rejection level of 2.5%.

Hitchhiking with and without Recombination: Genetic Distance from the Sweep

Figure 3a shows the power of the tests to detect a recent selective sweep ($T=0.001$, corresponding to roughly 400 years for *Drosophila*) as a function of the genetic distance from the selected site (ratio of *c* over *s*), for a given number of polymorphic sites remaining in the sample ($S=35$, and the expected value for $\theta=10$ in the absence of sweep). The left hand side of the figure corresponds to hitchhiking without recombination, for example, severe selective sweeps with fixation of the advantageous mutant before any recombination had occurred between the selected and the marker loci. Such events lead to star-like genealogies (Fig. 1c), an excess of rare variants and excess of haplotypes since most, if not all, polymorphism have arisen after the event. In this case, Tajima's *D*, Fu and Li's *D*, and Z_{nS} show a deficit of the statistics and tend to have higher power than the haplotype tests that show an excess of haplotypes. In this direction, haplotype tests and Z_{nS} are not conservative with respect to recombination and are thus more problematic. Of more interest is the detection of a deficit of haplotypes due to hitchhiking with recombination: a relatively "mild" sweep, with recombination events occurring between the selected locus and the marker during the selective stage. The peak obtained for intermediate distance from the selected site (Fig. 3a) corresponds to the deficit of haplotypes due to the survival of several lineages through the selective sweep stage (Fig. 1b). In this range of parameter values, haplotype tests, especially the *K* test, show high power, particularly when compared

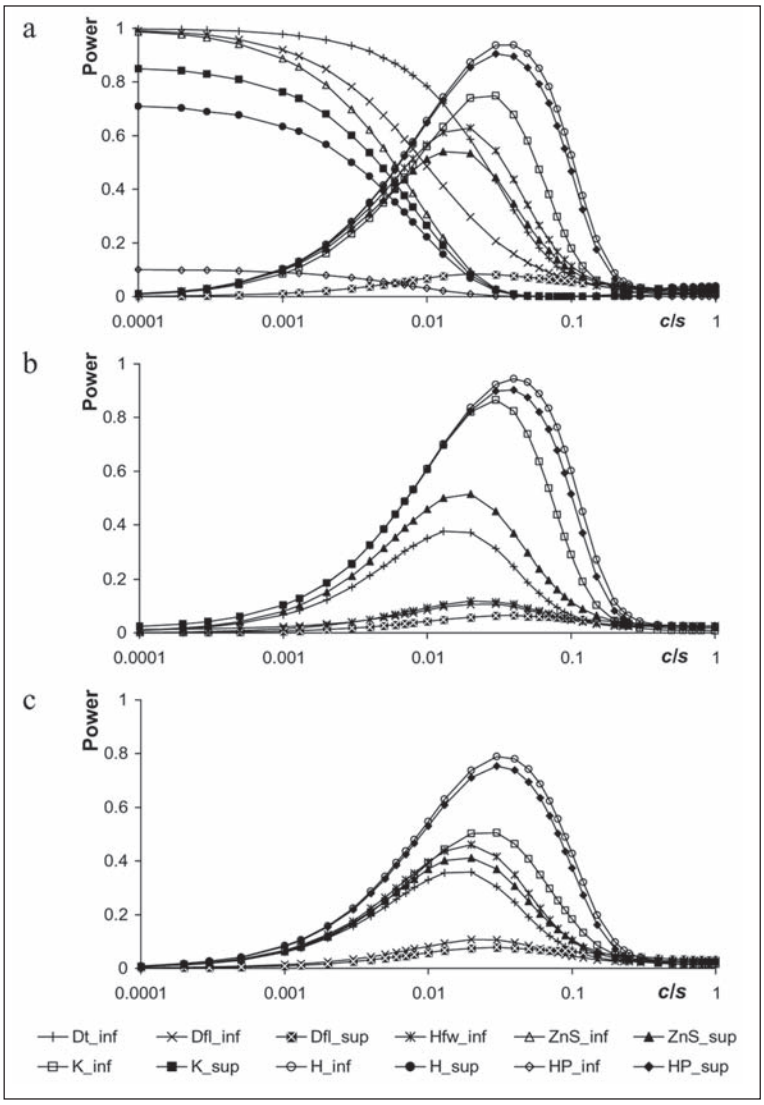

Figure 3. Power of several neutrality tests to detect a recent selective sweep (T=0.001 in $4N$ generations) as a function of the genetic distance from selected locus (c/s with α=10,000). n = 20, Nr = 0. Tests: *K*, *H*, *HP*, *Dt*: Tajima's *D; Dfl*: Fu and Li's *D*, *Hfw*: Fay and Wu's *H*; *ZnS*. Empty symbols: lower limit; Filled symbols: upper limit. a) All simulations are conditional on the number of segregating sites (S=35, close to the expected value for θ=10). Depending on the strength of the selective sweep, the deviation can be to one side or to the other. b) All simulations are conditional on the population mutation parameter (θ=10). The power to detect strong selective sweeps is substantially reduced due to the lack of variation that remains in the sample. c) The power simulations are conditional on θ=10, but the outcomes of the simulations are tested conditional on the resulting value of S. As compared to strict conditioning on θ, the power shows an overall reduction because it does not take into account the reduction in variation due to the selective sweep.

to frequency spectrum statistics such as Tajima's *D*. Interestingly, for an intermediate range of genetic distances, haplotype tests and Z_{nS} show substantial power in both directions: some genealogies show several lineages surviving the sweep while others do not, leading to opposite values of the statistics. [It is obviously the sum of the power in the two directions which is relevant if the tests were to be used in a bilateral way, e.g., on nonrecombining systems]. Put another way, their distribution is broadened. As a consequence, in multilocus studies, if the expected effect of selective sweeps uniformly depends on genetic distance from the closest selected site, the variance between loci can be drastically enhanced. Thus, while using the haplotype tests and Z_{nS}, finding departure from neutrality in opposite directions for different loci does not necessarily imply different parameter values of selective sweeps (such as different ratios of *c* over *s*, or different distances from the closest selected site). In contrast, frequency spectrum statistics show some power in a consistent direction, whatever the distance from the selected site.

Finally, conditioning the preceding results on a number of segregating sites *S* artificially levels off the power between different scenarios (Fig. 1b, 1c), because it imposes the same amount of genetic information on each of them.

Level of Variation and Distance from the Sweep

In practice, a population would have a given neutral mutation rate (a θ value). If we condition the tests on this particular θ value, the power virtually disappears for a recent severe selective sweep, simply because such event leaves no variation in the sample (compare Fig. 3b with 3a). An attempt to reach substantial power would then require extensive genotyping. For these reasons (together with the intragenic recombination issue), haplotype tests are mainly useful for detecting a deficit of haplotypes due to hitchhiking with recombination, i.e., a selective sweep in a DNA region distantly flanking the selected site. If the level of variation has been reduced and the real θ value is used to condition the tests, the power of Fay and Wu's test is drastically reduced (Fig. 3b). This seems to be due to the fact that this statistic, in contrast with other frequency spectrum statistics, is not normalized (the difference between the two estimators of θ is not divided by its variance). As a consequence, the variance and the width of the confidence interval for the of Fay and Wu's statistics are largely proportional to the level of variation. Conditioning of this test on *S*, following the procedure originally used by its authors, or using a normalized version of the statistics, should eliminate this effect.

In the preceding scheme, however, we conditioned all simulations on θ, assuming that it was known. In practice, this is not generally the case. In an attempt to mimic the conditions found in empirical studies, we used a more realistic procedure similar to that of ref. 43 (where it was used for other purposes: assessing the robustness of conditioning on *S*). We simulated the alternative hypothesis (selective sweep model) with a given θ value. For each simulated dataset resulting from such selective sweep, we tested it assuming that the θ value is unknown, and conditioned the tests on the observed number of mutations *S* that remain after the selective sweep – as we would do in practice on a real dataset. When the neutral model is simulated with a fixed θ and then tested according to the resulting *S*, this procedure rejects the neutral model in the proportion corresponding to the chosen threshold, as expected.[43] This is a computationally more demanding procedure as, for a given θ, a large variety of *S* values are obtained by simulations and we need the confidence intervals for all of them. We thus used only 5,000 simulations to compute the confidence interval for each possible *S* value (coalescent simulations are highly stochastic and a large number of iterations are needed to obtain precise estimates[14]). The resulting imprecision should be partly compensated by the large number of *S* values considered during the 100,000 power simulations. Such a procedure leads to an overall substantial

reduction in the power of the tests (compare Fig. 3c with 3b), because we do not use the information contained in the reduction of variation as compared to the expected level in the absence of selective sweep. For example, the number of haplotypes may be unexpectedly low given the number of mutations that remains after sweep, but it is then even less expected given the higher number of mutations we would anticipate if there were no sweep of variation.

Age of the Sweep

The above results are very similar to those obtained for a simple bottleneck model (unpublished results). In particular, the effects of the age of the sweep are quite trivial and closely match those of a bottleneck. Briefly, severe perturbations (hitchhiking without recombination or severe bottlenecks) can only be detected for an event of intermediate age, preferably using frequency spectrum statistics. During this age range, mutations have started to recover since the sweep, but the genealogy is still star-like. On the contrary, more moderate perturbations (hitchhiking with recombination or moderate bottlenecks) can be detected for recent events only (T<0.4 N), preferably using haplotype statistics (Fig. 4a). The reason for the rapid decay in power is that the signal of the sweep, in terms of haplotype structure, is provided by mutations present before the sweep, and mutations that appear after the sweep tend to obscure this signal. This contrasts with hitchhiking without recombination, where the only information is provided by the latter kind of mutations revealing the star-like genealogy.

Effect of the *s* Value for a Given c/*s*

For a given ratio of c over s, the effect of the sweep may still depend on the s value.[19,52,53] One obvious related effect is that the duration of the sweep ($-\ln(\varepsilon)/\alpha$) decreases with the increase in s. Hence, there is an age effect confounded with this s effect. Looking at the age of fixation of the sweep may not be relevant for different s values. Most of the coalescent events, and thus the bulk of the effect of sweep, occur at the beginning of the sweep when the advantageous allele is at low frequency, which leads to a high coalescence rate. Thus the time of occurrence of the advantageous mutation may be more relevant. We chose to keep this time constant and to allow s to vary, keeping c/s constant as well. [For comparison, note that the present procedure is in contrast with that of ref. 53 where it is the time of fixation that is set constant.] For a given age, our procedure sets up a lower bound for the possible range of s values for the sweep to be completed at the time of sampling. For the range of parameter values compatible with a quite recently completed sweep (0.04 N generations, when haplotype tests remain powerful; Fig. 4a), the s effect is rather weak (Fig. 5). The effect was undetectable for hitchhiking without recombination (results not shown). However, there may be additional effects for the more accurate stochastic treatments[19,52] (see concluding remarks).

Intragenic Recombination Effects for a Recent and Moderate Sweep

We used realistic values of intragenic recombination rate for the simulations under the selective sweep model ($4Nr$ =5 per locus, matching roughly 1kb in genomic regions with intermediate recombination rate in *Drosophila*). Computing confidence intervals using a model without intragenic recombination is conservative in the direction of an excess of haplotype structure (deficit of K, H, excess of HP, Z_{nS}), as recombination tends to increase the number of haplotypes. Frequency spectrum statistics are conservative in both directions in respect to intragenic recombination, as it does not affect the expectation of these statistics, and reduces the variance of all statistics.[34] Hence, such approach probably represents a safe procedure in practical situations.

As computed using this procedure, the effects of intragenic recombination on the power of the tests are minor (compare Fig. 6a with Fig. 3c), the Z_{nS} test being the most affected

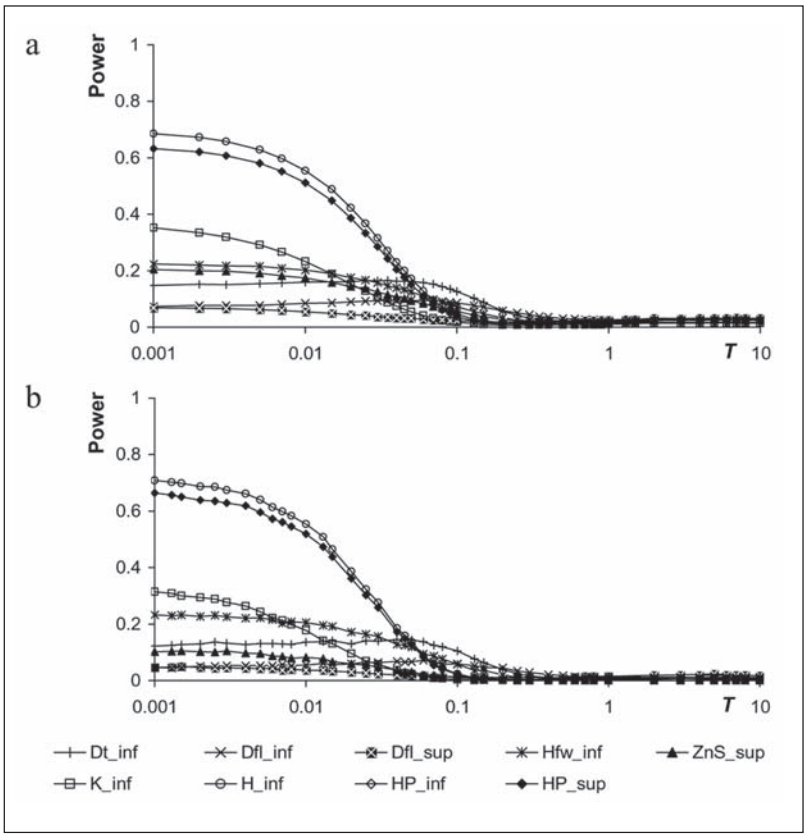

Figure 4. Power of several neutrality tests as a function of the age of a moderately distant sweep (C+ $4Nr$=600). Other parameters are same as for the (Fig. 3c). a) Without intragenic recombination ($4Nr$=0); b) With intragenic recombination ($4Nr$ =5; 10,000 runs per data point; curves for the directions that are nonconservative with respect to recombination are not shown).

(reduced power). Even if recombination is frequent before the sweep, only a few haplotypes survive the sweep through recombination into the advantageous background and increase in frequency thereafter. Such a moderate sweep has to be recent to be detected even in the absence of recombination as mutations that occur after the event tend to obscure the signature of the sweep (Fig. 4a). Therefore, recombination has little time to disrupt the haplotype structure during and after the sweep. This contrasts with the results of Wall[34] for a model of structured population, where there is plenty of time for recombination to disrupt the haplotype structure. In fact, in the presence of recombination the power of most tests is slightly increased. This may reflect the reduction of the variance of the statistics in the presence of recombination. For moderately distant sweeps with recombination, there is no noticeable acceleration of the decay in power with the age of the event (compare Fig. 4b with Fig. 4a). [Note that the order in which the curves with and without recombination cross the Y axes depends largely on the genetic distance considered.] The accelerated decay effect can be substantial only if the Nr value is of a higher order of magnitude than θ (results not shown), which seems to be a rare situation in biological systems.

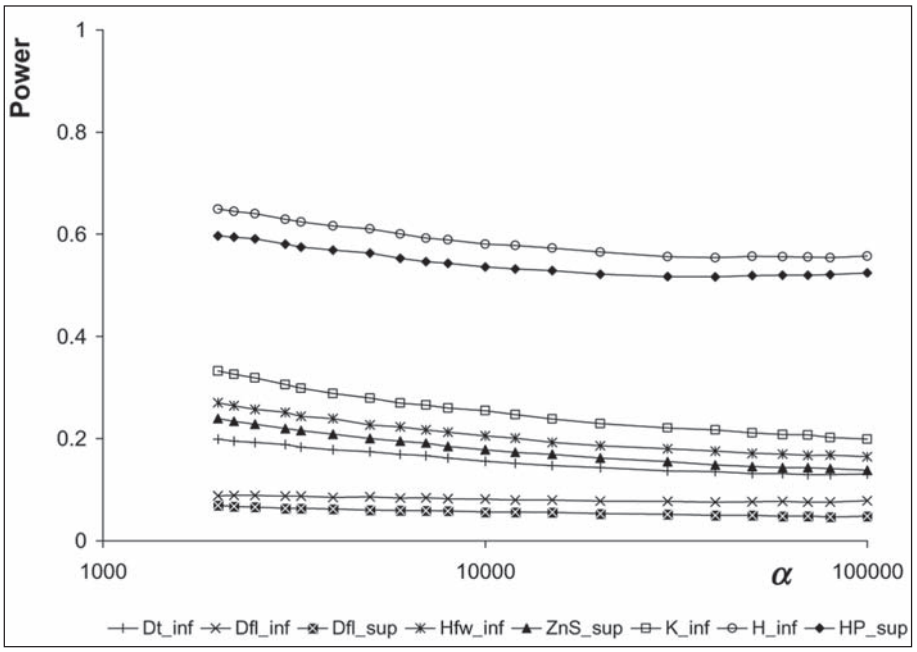

Figure 5. Power of the tests as a function of the selection coefficient for a fixed c/s (0.06) and a given age of occurrence of the advantageous mutation (T_i=0.01). Other parameters are same as for Figure 3c.

On the other hand, if we knew the exact recombination rate parameter of the population, we could use coalescent simulations with recombination and condition the tests on this recombination rate value. Conditioning on this parameter values increases the power of the tests as it diminishes the probability of observing so few haplotypes and reduces the variance of all statistics (Fig. 6b to be compared with Fig. 6a). This is probably an overestimate of the gain that could be obtained in practice, as this parameter value would show some uncertainty, and a lower bound of this parameter should be used for the test to remain conservative.

Concluding Remarks

In the light of the above difficulties, it is advisable to be cautious in applying haplotype tests and interpreting the results they provide. In particular, this concerns dealing with intragenic recombination, and when considering the deviation from neutrality towards deficit or excess of statistics since behavior of many tests is asymmetric. Several alternative hypotheses should be considered and the tests should be used together with other statistics to provide complementary insights. This, however, raises the question of multiple testing.

The issue of distinguishing between bottlenecks and selective sweep type of explanations is discussed more specifically elsewhere.[54] Briefly, considering a single locus, the distribution of haplotypes differs between the two processes for moderate perturbations.[19] During hitchhiking with recombination, one peculiar lineage, originally carrying the advantageous mutations, increases in frequency more drastically than other lineages (Fig. 1b). As a result, selective sweep predicts one major family of lineages and a few rare variants, thereby increasing the power of H and HP as compared to the bottleneck case. However, the haplotype frequency effects could arise from many different phenomena and it may be unwise to rely exclusively on such an

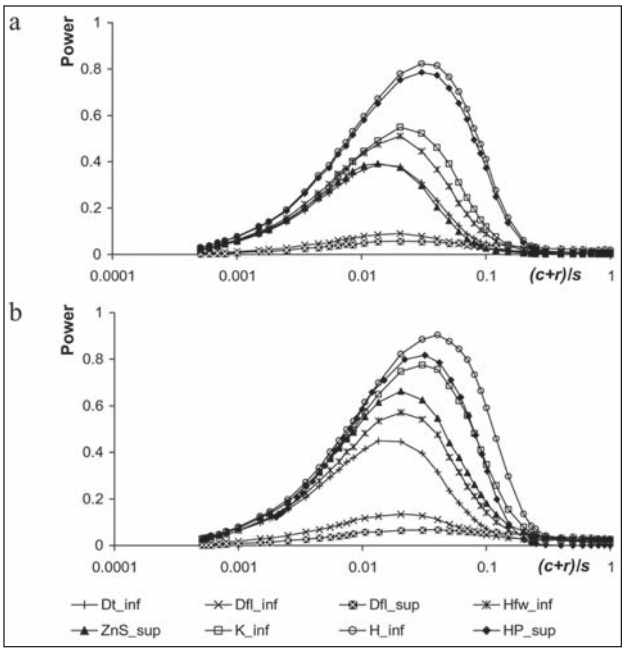

Figure 6. Power of several neutrality tests in the presence of intragenic recombination ($4Nr$ =5) as a function of the genetic distance from the selected site (actually, the intragenic recombination rate is included in this distance, which sets a lower bound for the minimum X value used). 10,000 runs were used for each data point in the power simulations. Other parameters are same as for the (Fig. 3c). Curves for the directions that are nonconservative with respect to recombination are not shown. a) Tested assuming no recombination. b) Tested conditioning on the actual recombination rate.

approach. The most obvious way to distinguish hitchhiking and bottlenecks is through multilocus comparisons: unlike bottlenecks, which are expected to affect all loci to a similar extent, selective sweeps are expected to affect different loci to various degrees, depending on their level of linkage with the selected loci. This is the basis of the HKA neutrality test[37] which compares the ratio of polymorphism to divergence for several loci and thus in principle distinguishes between the selective effects and the alternative demographic explanations such as bottlenecks, provided that the sampling schemes for the different loci are consistent. While this type of evidence appears intuitively reasonable, the neutral assumption of a panmictic population of constant size may not be conservative with regard to bottlenecks or population structure effects, which are highly stochastic and may produce very different patterns on various loci.[54] It may thus be necessary to explicitly implement the models of the bottleneck and of the selective sweep and to compare the two with, for example, likelihood ratio tests.[38]

The lineage effect described above for a given locus also impacts the frequency spectrum of mutations, which shifts towards an excess of rare mutations, and especially those of ancestral origin, under hitchhiking with recombination (Fig. 1b). This property is used in the neutrality test proposed by Fay and Wu,[22] with assumption that the ancestral state of variants can be inferred from an outgroup because the proportion of homoplasies at this interspecific level is considered low as compared to the proportion of rare—presumably ancestral—variants (which itself remains generally low). The expected proportion of homoplasies is usually computed assuming a simple mutational model, without codon bias, transition-transversion bias or het-

erogeneity of the neutral mutation rate along the sequence. As these assumptions are not conservative, when the null hypothesis is rejected, potential alternative explanations include mutational effects.[55]

The hitchhiking model that we use also shows some limitations. The change in frequency of the advantageous mutations is modeled with a deterministic approximation. In a population of finite size, allele frequency trajectories are highly stochastic when the frequency is close to the extreme values (soon after the occurrence of the advantageous mutation and when it is close to fixation). These stochastic periods represent a substantial fraction of the duration of selective stage. Such drift effects tend to shorten the selective stage and, therefore, to enhance the hitchhiking effect. Most advantageous variants would disappear by drift soon after their occurrence, and the variants that would proceed to fixation tend to be the ones which, by chance, increase more rapidly in frequency in their early stages. When at high frequency, an advantageous mutant goes to fixation earlier than predicted, simply because the fixation is an absorbing boundary. As for the genealogy of n genes, we expect these effects to be small.[53] These ranges of frequency are similar to the neutral stage: most lineages are in the common genetic background (e.g., in the neutral background when the advantageous mutant is still at low frequency); their probability of common ancestry is thus little affected and their probability of transition from one background to the other is low (proportional to x in the above case).

Finally, while surveys addressing the properties of statistics provide useful qualitative and quantitative clues, they are generally biased in several ways, especially when derived from simulation studies. It is not clear to what extent the models chosen and the set of parameter values used are relevant or primarily reflect the expectations of the authors. Furthermore, the same authors often propose a statistics and also assess its properties (the present chapter included!), thus raising the question of nonindependence of the survey.

In conclusion, haplotype statistics can show high power in various circumstances, including hitchhiking effects, especially in the case of incomplete hitchhiking, and they may complement other kinds of information. Perhaps the best evidence of the relevance of haplotype statistics is the frequency of their use on actual datasets.[17,18,24,29,46,47,51,56-58]

Acknowledgements

We thank M. Cobb, N. Galtier, A. Navarro and D. Nurminsky for comments on previous version of the manuscript. A user-friendly version of the haplotype tests is available at http://www.snv.jussieu.fr/mousset/. Our research is supported by CNRS UMR 7625, GDR 1928 and the Ecole Pratique des Hautes Etudes.

References

1. Kimura M. The neutral theory of molecular evolution. Sci Am 1979; 241:98-100.
2. Moriyama EN, Powell JR. Intraspecific nuclear DNA variation in Drosophila. Mol Biol Evol 1996; 13:261-277.
3. Barton NH. Genetic hitchhiking. Philos Trans R Soc Lond B Biol Sci 2000; 355:1553-1562.
4. Begun DJ, Aquadro CF. Molecular population genetics of the distal portion of the X chromosome in Drosophila: Evidence for genetic hitchhiking of the yellow-achaete region. Genetics 1991; 129:1147-1158.
5. Begun DJ, Aquadro CF. Levels of naturally occuring DNA polymorphism correlate with recombination rates in D. melanogaster. Nature 1992; 356:519-520.
6. Berry AJ, Ajioka JW, Kreitman M. Lack of polymorphism on the Drosophila fourth chromosome resulting from selection. Genetics 1991; 126:1111-1117.
7. Nachman MW, Bauer VL, Cromwell SL et al. DNA variability and recombination rate at X-linked loci in Humans. Genetics 1998; 150:1133-1141.
8. Nachman MW. Patterns of DNA variability at X-linked loci in Mus domesticus. Genetics 1997; 147:1303-1316.

9. Baudry E, Kerdelhue C, Innan H et al. Species and recombination effects on DNA variability in the tomato genus. Genetics 2001; 158:1725-1735.
10. Marais G, Mouchiroud D, Duret L. Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. Proc Natl Acad Sci USA 2001; 98:5688-5692.
11. Charlesworth B, Morgan MT, Charlesworth D. The effect of deleterious alleles on neutral molecular variation. Genetics 1993; 134:1289-1303.
12. Hudson RR, Kaplan NL. Deleterious Background selection with recombination. Genetics 1995; 141:1605-1617.
13. Braverman JM, Hudson RR, Kaplan NL et al. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics 1995; 140:783-796.
14. Hudson RR. The how and why of generating gene genealogies in: Takahata and Clarck, eds. Mecanism of molecular evolution. Japan scientific societies press Sinauer associates, inc, 1993:23-36.
15. Begun DJ, Aquadro CF. Evolution at the tip and base of the X chromosome in an African population of Drosophila melanogaster. Mol Biol Evol 1995; 12:382-390.
16. Carr M, Soloway JR, Robinson TE et al. An investigation of the cause of low variability on the fourth chromosome of Drosophila melanogaster. Mol Biol Evol 2001; 18:2260-2269.
17. Wang W, Thornton K, Berry A et al. Nucleotide variation along the Drosophila melanogaster fourth chromosome. Science 2002; 295:134-137.
18. Kirby DA, Stephan W. Multi-locus selection and the structure of variation at the white gene of Drosophila melanogaster. Genetics 1996; 144:635-645.
19. Barton NH. The effect of hitch-hiking on neutral genealogies. Genet Res 1998; 72:123-133.
20. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 1989; 123:585-595.
21. Fu YX, Li WH. Statistical tests of neutrality of mutations. Genetics 1993; 133:693-709.
22. Fay JC, Wu CI. Hitchhiking under positive Darwinian selection. Genetics 2000; 155:1405-1413.
23. Kim Y, Stephan W. Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 2002; 160:765-777.
24. Hudson RR, Bailey K, Skarecky D et al. Evidence for positive selection in the superoxide dismutase (Sod) region of Drosophila melanogaster. Genetics 1994; 136:1329-1340.
25. Strobeck C. Average number of nucleotide differences in a sample from a single subpopulation a test for population subdivision. Genetics 1987; 117:149-154.
26. Fu YX. New statistical tests of neutrality for DNA samples from a population. Genetics 1996; 143:557-570.
27. Fu YX. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics 1997; 147:915-925.
28. Depaulis F, Veuille M. Neutrality tests based on the distribution of haplotypes under an infinite-site model. Mol Biol Evol 1998; 15:1788-1790.
29. Andolfatto P, Wall JD, Kreitman M. Unusual haplotype structure at the proximal breakpoint of In(2L)t in a natural population of Drosophila melanogaster. Genetics 1999; 153:1297-1311.
30. Griffiths RC. The number of alleles and segregating sites in a sample from the infinite-alleles model. Adv Appl Prob 1982; 14:225-239.
31. Nei M. Molecular evolutionary genetics. New York: Columbia University Press, 1987.
32. Watterson GA. The homozygosity test of neutrality. Genetics 1978; 88:405-417.
33. Kelly JK. A test of Neutrality based on interlocus associations. Genetics 1997; 146:1197-1206.
34. Wall JD. Recombination and the power of statistical tests of neutrality. Genet Res 1999; 74:65-69.
35. Fay J, Wu C. A human population bottleneck can account for the discordance between patterns of mitochondrial versus nuclear DNA variation. Mol Biol Evol 1999; 16:1003-1005.
36. Takahata N, Nei M. Allelic Genealogy under overdominant and frequency-dependent selection and polymorphism of Major Histocompatibility Complex loci. Genetics 1990; 124:967-978.
37. Hudson RR, Kreitman M, Aguadé M. A test of neutral molecular evolution based on nucleotide data. Genetics 1987; 116:153-159.
38. Galtier N, Depaulis F, Barton NH. Detecting bottlenecks and selective sweeps from DNA sequence polymorphism. Genetics 2000; 155:981-987.
39. Ewens WJ. The sampling theory of selectively neutral alleles. Theor Popul Biol 1972; 3:87-112.
40. Watterson GA. On the number of segregation sites. Theor Popul Biol 1975; 7:256-276.

41. Markovtsova L, Marjoram P, Tavaré S. On a test of Depaulis and Veuille. Mol Biol Evol 2001; 18:1132-1133.
42. Depaulis F, Mousset S, Veuille M. Haplotype tests using coalescent simulations conditional on the number of segregating sites. Mol Biol Evol 2001; 18:1136-1138.
43. Wall JD, Hudson RR. Coalescent simulations and statistical tests of neutrality. Mol Biol Evol 2001; 18:1134-1135.
44. Wall JD. A comparison of estimators of the population recombination rate. Mol Biol Evol 2000; 17:156-163.
45. Hudson RR, Kaplan NL. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 1985; 111:147-164.
46. Bénassi V, Depaulis F, Meghlaoui GK et al. Partial sweeping of variation at the Fbp2 locus in a west African population of Drosophila melanogaster. Mol Biol Evol 1999; 16:347-353.
47. Depaulis F, Brazier L, Veuille M. Selective sweep at the Drosophila melanogaster suppressor of hairless locus and its association with the In(2L)t inversion polymorphism. Genetics 1999; 152:1017-1931.
48. Przeworski M, Wall JD. Why is there so little intragenic linkage disequilibrium in humans? Genet Res 2001; 77:143-151.
49. Przeworski M, Wall JD, Andolfatto P. Recombination and the frequency spectrum in Drosophila melanogaster and Drosophila simulans. Mol Biol Evol 2001; 18:291-298.
50. Andolfatto P, Przeworski M. A genome-wide departure from the standard neutral model in natural populations of drosophila. Genetics 2000; 156:257-268.
51. Kirby DA, Stephan W. Haplotype test reveals departure from neutrality in a segment of the white gene of Drosophila melanogaster. Genetics 1995; 141:1483-1490.
52. Stephan W, Wiehe THE, Lenz MW. The effect of strongly selected substitutions on neutral polymorphism: Analytical results based on diffusion theory. Theor Popul Biol 1992; 41:237-254.
53. Przeworski M. The signature of positive selection at randomly chosen Loci. Genetics 2002; 160:1179-89.
54. Depaulis F, Mousset S, Veuille M. Power of neutrality tests to detect bottlenecks and hitchhiking. J Mol Evol 2003; 57(Suppl1):S190-200.
55. Baudry E, Depaulis F. Effect of misoriented sites on neutrality tests with outgroup. Genetics 2003; 165:1619-1622.
56. Cirera S, Aguadé M. Evolutionnary history of the sex-peptide (Acp70A) gene region in Drosophila melanogaster. Genetics 1997; 147:189-197.
57. Hamblin MT, Veuille M. Population structure among African and derived populations of Drosophila simulans: Evidence for ancient subdivision and recent admixture. Genetics 1999; 153:305-317.
58. Andolfatto P, Kreitman M. Molecular variation at the In(2L)t proximal breakpoint site in natural populations of Drosophila melanogaster and D. simulans. Genetics 2000; 154:1681-1691.

Chapter 5

A Novel Test Statistic for the Identification of Local Selective Sweeps Based on Microsatellite Gene Diversity

Christian Schlötterer and Daniel Dieringer

Genome wide population surveys have recently been established as a promising approach for the identification of genomic regions subject to directional selection. Nevertheless, the analysis of multiple markers requires novel approaches for the identification of selection. In this report we introduce a new test statistic, lnRH, which is based on the relative gene diversity in two populations. Similar to the previously introduced lnRV test statistic, the distribution of lnRH captures the demographic history of the populations as well as variation in microsatellite mutation rates among loci. Using coalescent based computer simulations we demonstrate that the lnRH test statistic has a higher power than lnRV. Secondly, using lnRH and lnRV jointly, we show that the number of false positives (type I error) can be reduced by a factor of three.

The question of how populations adapt to their environment has been a long-standing dispute in biology. After numerous allozyme and chromosomal inversion studies, the advance of molecular biology allowed the analysis of candidate genes.[1] Although this approach has greatly advanced our current understanding of selection, it has certain limitations for the characterization of adaptation events. Probably the greatest shortcoming of such a candidate gene approach is its dependence on a priori information about possible candidate genes. Given our still limited understanding of the molecular basis of adaptation, other approaches are required.

Recent technological advances have provided the opportunity to expand the analysis of single loci to genome wide surveys. One of these approaches focuses on the transcriptome, the entire set of RNAs in an organism. By comparing the expression level in individuals from different populations, it should be possible to identify a set of candidate genes, which may be involved in adaptation. This approach has been successfully used for evolved yeast strains[2] and *D. melanogaster* lines, which have been subjected to strong selection for positive and negative geotaxis.[3] In addition to technical difficulties posed by tissue- and development-specific expression alterations and by the requirement of large expression differences for successful detection, this approach suffers from shortcomings that were also noted for phenotypic characters.[4]

Alternatively, adaptation could be studied by exploiting general principles of population genetic: (1) unless lost by genetic drift, beneficial mutations are expected to spread through a population until they become eventually fixed; (2) such a spread of a beneficial mutation leaves characteristic traces at the selected site and its flanking region.[5-7] Hence, genome scans surveying

Selective Sweep, edited by Dmitry Nurminsky. ©2005 Eurekah.com and Kluwer Academic/Plenum Publishers.

patterns of variability in natural populations adapted to different environments could serve as a tool for the identification of genomic regions bearing a beneficial mutation. While the molecular tools are well-developed to perform such genome scans, only a few test statistics have been developed for this purpose (reviewed in ref. 8).

Their high polymorphism, and straightforward and cost-effective analysis made microsatellites the markers of choice for a recent series of genome-wide scans for genomic regions bearing a beneficial mutation.[6,9-14] Nevertheless, the interpretation of microsatellite polymorphism is often complicated by a locus specific mutation rate.[15,16] This heterogeneity in mutation rate can be accounted for by analyzing the ratio of the observed variance in repeat number at each locus.[14] As all loci have the same expectation, irrespective of their mutation rate, it is possible to identify loci which differ significantly from the remainder of the genome. This lnRV test statistics suffers from two disadvantages: (1) the variance in repeat number has a very large variance, which reduces the power to detect selected loci; (2) depending on the α-level and on the number of loci analyzed, the number of false positives can be quite large.

Here, we extend the approach of Schlötterer[14] by using gene diversity rather than variance in repeat number as an estimator of variability. We show that lnRH has a higher power to identify selected loci than lnRV and that the joint consideration of lnRH and lnRV reduces the number of false positives about three-fold.

Material and Methods

Computer Simulations

We used standard coalescent simulations[17] to simulate the allele distribution in two populations. The original C code was modified to account for the stepwise mutation process of microsatellites. After simulating the number of mutations occurring on a given branch, they were converted into microsatellite mutations by adding or removing (with equal probabilities) one repeat unit for each mutation. All simulated loci were assumed to be independent (unlinked). If not designated otherwise, we made the standard assumptions of the coalescent process, such as neutrality, constant population sizes and panmixia.

Between 100 and 10,000 loci were simulated for two independent populations using the unbiased stepwise mutation model.[18] For each locus the lnRH and lnRV test statistic was calculated. When variation in microsatellite mutation rate was incorporated in the simulations, the mutation rates varied by a factor of 10 drawn from a uniform distribution. For those simulations, mean θ-values are reported (θ=$4N_e\mu$). For a subset of simulations we used the more general two-phase mutation model[19] rather than the strict stepwise mutation model. The two-phase mutation model assumes that a certain fraction of mutations encompasses multiple repeat units. The size change for such mutations was drawn from a uniform distribution ranging from 1 to a specified maximum.

We simulated two demographic models, population bottleneck and population expansion, as outlined in Hudson[17] by assuming an instantaneous change in population size. The demographic model is specified by the factor f by which the population size changes and the time t (in $4N_e$ generations) when the demographic event occurs.

To determine the power of the lnRV and lnRH test statistics, we modified the neutral coalescent simulations. 100 loci were simulated for each parameter set and one of the loci was assumed to be linked to a genomic region subjected to a selective sweep. Simulations of this locus were based on a reduction in population size at a given time interval. The intensity of the selection and distance to the selected site were jointly considered by specifying the magnitude of the reduction in population size. The remaining loci were simulated under the standard coalescent model.

Test Statistics

Variability at a given microsatellite locus could be either measured by the variance in repeat number ($V=2N_e\mu$) or gene diversity ($H=1-(1)/(\sqrt{1+8N_e\mu})$. Schlötterer[14] used the variance in repeat number to design a test statistic lnRV

$$\ln\left[\mathrm{E}\left(RV\right)\right] = \ln\left[\mathrm{E}\left[\left(\frac{\frac{1}{2}\theta_{Pop1}}{\frac{1}{2}\theta_{Pop2}}\right)\right]\right] \cong \ln\left[\mathrm{E}\left[\left(\frac{\left(2N_{e_{Pop1}}\mu\right)}{\left(2N_{e_{Pop2}}\mu\right)}\right)\right]\right] = \left[\frac{\mathrm{E}\left(V_{Pop1}\right)}{\mathrm{E}\left(V_{Pop2}\right)}\right] \tag{1}$$

The corresponding test statistic for gene diversity lnRH is calculated as:

$$\ln\left[E\left(RH\right)\right] = \ln\left[E\left(\frac{\theta_{Pop1}}{\theta_{Pop2}}\right)\right] \cong \ln\left[\mathrm{E}\left[\frac{\left(\left(\frac{1}{1-H_{Pop1}}\right)^2-1\right)\frac{1}{8\mu}}{\left(\left(\frac{1}{1-H_{Pop2}}\right)^2-1\right)\frac{1}{8\mu}}\right]\right] = \ln\left[\frac{\mathrm{E}\left[\left(\frac{1}{1-H_{Pop1}}\right)^2-1\right]}{\mathrm{E}\left[\left(\frac{1}{1-H_{Pop2}}\right)^2-1\right]}\right] \tag{2}$$

It has to be noted that equation 1 and 2 are only approximations, and the delta method[20] performs better. Nevertheless, Schlötterer[14] showed that the differences are minor, thus we used equation 1 and 2.

Some simulations resulted in monomorphic samples, in particular for low θ-values. In these cases neither lnRV nor lnRH would be defined, therefore we substituted one allele in the study for an allele which differed by one repeat unit. We have chosen this approach as the most conservative correction possible. Furthermore, it has the advantage that it also accounts for differences in sample size. Alternatively, a small value could be added to monomorphic samples, but the choice of the value is rather arbitrary—and thus also is an associated significance level. For this study we therefore preferred the first approach.

Both test statistics lnRV and lnRH can be assumed to be independent of the mutation rate of the microsatellite, and all loci have the same expectation for each of the two statistics. Nevertheless, genetic drift results in some variation of coalescent times among the loci studied. Schlötterer[14] showed that for a wide range of parameters the distribution of lnRV values could be approximated by a Gaussian distribution. To evaluate the shape of the lnRH distribution we followed the outline of Schlötterer.[14] first a nonparametric Kolmogorov-Smirnov test was used to evaluate the distribution of 1,000 simulated lnRH values. Second, a "tail" test was performed on the same data set (see Schlötterer ref. 14 for details).

Results

Verification of the lnRH Test Statistic

In order to use the lnRH test statistic analogous to lnRV, it is important that the lnRH values are normally distributed. We used standard coalescent simulations to determine the distribution of lnRH values under a range of evolutionary scenarios.

Dependence on Mutation Rate and Model

We performed computer simulations using a broad range of θ-values and found that the distribution of lnRH values was well approximated by a normal distribution. In addition, when different θ-values were assumed for both populations, no deviation from a normal distribution was detected (Table 1). As microsatellite mutation rates have been found to differ substantially among loci,[15,16] we also tested whether variation in mutation rate affects the normal distribution of lnRH. To account for heterogeneity in mutation rates, θ-values were drawn from a uniform distribution resulting in an up to 10-fold increase in mutation rate, hence $\bar{\theta}$ is

Table 1. Variance of lnRH and lnRV for different θ-values based on computer simulations of 10,000 loci in two neutrally evolving populations

	θ=3	θ=5	θ=10	θ=50	$\bar{\theta}$=3	$\bar{\theta}$=30	$\bar{\theta}_1/\bar{\theta}_2$=5/500	$\bar{\theta}_1/\bar{\theta}_2$=500/5
lnRH	1.15*	0.90*	0.70*	0.49*	1.64*	0.56*	0.60*	0.61*
lnRV	1.53*	1.48*	1.39*	1.34*	1.82*	1.38*	1.36*	1.39*

* No significant deviations from normal distribution by tail test (P> 0.2) and Kolmogorov-Smirnov test (P> 0.3)

the expectation of the θ-values used for the simulations. No significant deviation from a normal distribution was detected (Table 1).

The simplest model of microsatellite evolution assumes that gains and losses of single repeat units occur with the same frequency. Experimental evidence, however, suggests that microsatellite mutations could encompass more than a single repeat unit (two-phase model). We investigated the influence of a more general microsatellite mutation model on the distribution of lnRH values. Mutation step sizes were drawn from a uniform distribution. For a broad range of parameters, no deviation from a normal distribution of lnRH values could be detected (Table 2).

Demography

We used two simple models of demography to test the behavior of the lnRH test statistic: a recent bottleneck and population expansion in one of the two populations, while the other population remained at constant size. For the simulation runs that assumed recent bottlenecks and a low mutation rate, we found a significant deviation from a normal distribution (Table 3). This deviation can be attributed to a large number of invariant loci. Note that for lnRV no significant deviation from a normal distribution was observed, even for recent bottlenecks, while Schlötterer (2002) noted a deviation from a normal distribution. This apparent discrepancy results from the different treatment of monomorphic loci. Irrespective of this treatment, both lnRH and lnRV in general should not be applied to data sets that contain a large number of monomorphic loci.

Table 2. Variance of lnRH and lnRV for different θ-values based 10,000 loci in two neutrally evolving populations

	K = 0		K = 0.2		K = 0.4	
	lnRH	lnRV	lnRH	lnRV	lnRH	lnRV
S = 5	0.90*	1.48*	0.95*	1.69*	0.86*	1.60*
S = 10			0.93*	2.36*	0.99*	2.42*

K is the probability of mutation encompassing more than one repeat unit, and S is the upper boundary of size change by a single mutation event. θ = 5 (in both populations). * No significant deviations from normal distribution by tail test (P> 0.2) and Kolmogorov-Smirnov test (P> 0.3)

Table 3. Variance of lnRH and lnRV when one population had passed through a bottleneck

	θ=3		θ=5		θ=10		θ=50	
	lnRH	**lnRV**	**lnRH**	**lnRV**	**lnRH**	**lnRV**	**lnRH**	**lnRV**
No bottleneck	1.15	1.53	0.90	1.48	0.70	1.39	0.49	1.34
t=0.1	1.74	1.48	1.14	1.17	0.75	1.03	0.43	0.94
t=0.05	2.10	1.74	1.43**	1.39	0.91	1.15	0.47	1.02
t=0.01	3.15***	2.59	2.44****	2.16	1.40	1.56	0.58	1.24

A total of 10,000 microsatellite loci were simulated for two populations and *f* (factor by which the population expanded) was set to 0.1. *t*= time (in $2N_e$) elapsed since the bottleneck. Significant deviations from normal distribution: ** P< 0.05 tail test; *** P< 0.05 Kolmogorov-Smirnov test; **** P< 0.05 Kolmogorov-Smirnov test and tail test

Some simulations of a population expansion (Table 4) also failed the tail-test. A closer inspection of the distribution of lnRH values indicated a systematic surplus of positive lnRH values, while fewer than expected negative lnRH values were observed (data not shown). This suggests that the lnRH test statistic is conservative for the identification of selective sweeps in the expanded population, but not for the population which remained at constant size. Despite being not statistically significant, lnRV shows the opposite trend (in particular for large θ-values, data not shown).

Comparison between lnRV and lnRH

As shown above, lnRH follows a normal distribution over a wide range of parameters-similar to lnRV. Hence, the interesting question is, which test statistic is better suited for the identification of loci linked to a selected site. As loci subjected to a selective sweep are expected to be located outside of the distribution of the remaining neutrally evolving loci, the test will have more power if the variance of the test statistic is small.

Table 4. Variance of lnRH and lnRV with one recently expanded population

	θ=3		θ=5		θ=10		θ=50	
	lnRH	**lnRV**	**lnRH**	**lnRV**	**lnRH**	**lnRV**	**lnRH**	**lnRV**
neutral	1.15	1.53	0.90	1.48	0.70	1.39	0.49	1.34
t=0.1, *f*=10	1.04**	1.38	0.81**	1.30	0.63	1.22	0.47**	1.21
t=0.1, *f*=100	0.75**	0.99	0.59	0.91	0.46	0.86**	0.59	0.91**
t=0.01, *f*=10	0.75	0.96	0.60	0.91	0.48	0.88	0.35	0.85
t=0.01, *f*=100	0.62	0.81	0.49	0.75	0.38	0.73	0.27	0.71

A total of 10,000 microsatellite loci were simulated for two populations. *t*= time (in $2N_e$) since the expansion; *f*= factor by which the population expanded. Significant deviations from normal distribution: ** P< 0.05 tail test. No deviation from normality was detected by the Kolmogorov-Smirnov test.

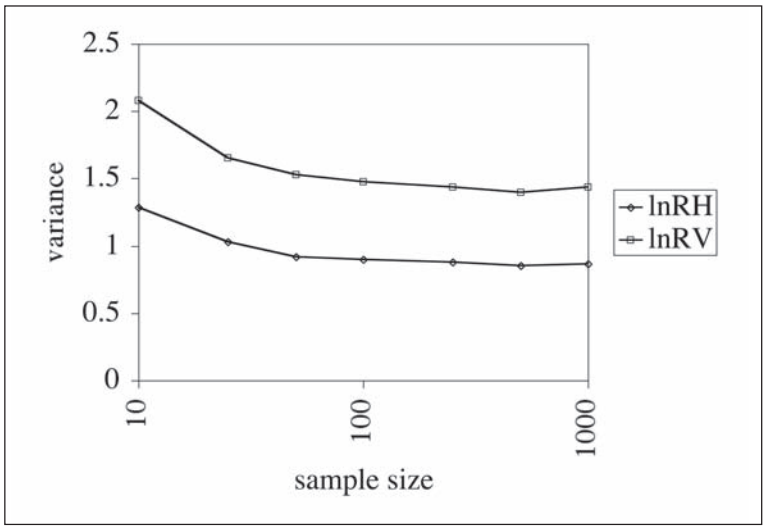

Figure 1. Influence of the sample size (in chromosomes) on the standard deviation of the lnRH (◇) and lnRV (□) test statistic. Variances were measured on 10,000 independently simulated microsatellite loci with θ =5 in both populations

Influence of Sample Size

The variance of lnRV and lnRH was determined for 10,000 microsatellite loci using different sample sizes. While both test statistics had a larger variance for small sample sizes, lnRH consistently had a smaller variance than lnRV (Fig. 1). Note that our preset condition of the presence of at least two different chromosomes in a sample has a particularly pronounced effect for small sample sizes: the minimal variance in repeat number in a sample of 10 is 0.1, while in a sample of 100 it is 0.01.

Statistical Power

For a direct comparison of the statistical power of lnRH and lnRV, we simulated 100 microsatellite loci, of which 99 evolved neutrally and one was assumed to be linked to a selected site. For each of the test statistics, the θ-values used in the computer simulations had only a very limited influence on the power of the tests (Table 5). Consistent with previous studies[14,21] we found that the power to detect older sweeps was lower, and that sweeps resulting in a stronger reduction in effective population size were easier to detect (Table 5). Interestingly, irrespective of the parameters used for the computer simulations, the lnRH test had a significantly higher power. The superiority of the lnRH test statistic became even more apparent when a more general mutation model was considered. While the power of the lnRV test statistic decreased under the two-phase model, almost no difference in power was noted for the lnRH test statistic (Table 6).

Joint Analysis of lnRH and lnRV

Gene diversity and variance are two different estimators of variability. As the amount of variability in a population sample is governed by the underlying genealogical structure, one may expect gene diversity and variance to be highly correlated. To test this, we determined the correlation between lnRH and lnRV using a set of neutrally evolved populations. Table 7 indicates that the two test statistics are not very strongly correlated and only about 70% of the

Table 5. Statistical power of the lnRH and lnRV test statistic measured by the fraction of correctly inferred selected loci

	θ =3		θ =5		θ =10		θ =50	
	lnRH	**lnRV**	**lnRH**	**lnRV**	**lnRH**	**lnRV**	**lnRH**	**lnRV**
t_S=0.2, f_S=0.01	0.37	0.18	0.32	0.16	0.26	0.16	0.22	0.16
t_S=0.1, f_S=0.01	0.52	0.35	0.54	0.31	0.51	0.34	0.43	0.34
t_S=0.05, f_S=0.01	0.77	0.56	0.82	0.59	0.82	0.57	0.79	0.57
t_S=0.05, f_S=0.001	0.8	0.61	0.87	0.62	0.9	0.62	0.85	0.66

One locus was subjected to directional selection and 99 loci evolved neutrally. 1000 simulation runs were performed. t_S= time point when selection occurred; f_S= factor by which variability was reduced at the selected locus

Table 6. Statistical power of the lnRH and lnRV test statistic for the two phase model (TPM)

	***K* = 0**		***K* = 0.2**		***K* = 0.4**	
	lnRH	**lnRV**	**lnRH**	**lnRV**	**lnRH**	**lnRV**
S = 5	0.82	0.59	0.80	0.55	0.82	0.52
S = 10			0.82	0.43	0.79	0.41

One locus was subjected to directional selection and 99 loci evolved neutrally. 1000 simulation runs were performed. f_S = 0.01. t_S = 0.05 (see table 5). See table 2 for an explanation of *K* and *S*. θ = 5 (in both populations).

Table 7. Correlation (r) between lnRH and lnRV

	θ =3	θ =5	θ =10	θ =50
neutral	0.753	0.726	0.706	0.724
bottleneck (0.1, 0.05)	0.875	0.816	0.772	0.766
expansion (0.1, 10)	0.782	0.766	0.759	0.752
TPM (0.2, 5)	n.d.	0.710	n.d.	n.d.
TPM (0.2, 10)	n.d.	0.603	n.d.	n.d.

n.d.= not determined, for simulations assuming a demographic change values in brackets indicate t_S and f_S (see Table 5), TPM= two phase model, values in brackets are *K* and *S* (see Table 2)

Table 8. Fraction of loci which are significant for both test statistics lnRH and lnRV

θ=3	θ=5	θ=10	θ=50	$\bar{\theta}_1/\bar{\theta}_2$=5/500	$\bar{\theta}_1/\bar{\theta}_2$=500/5	θ=5 TPMa	θ=5 TPMb
1.8%	1.4%	1.5%	1.6%	1.4%	1.5%	1.6%	1.8%

A total of 10,000 microsatellite loci were simulated for two populations. TPMa= two phase model with K=0.2, S=5, TPMb= two phase mutation model with *K*=0.4, *S*=10 (see Table 2)

variation could be explained by the correlation between lnRH and lnRV. This suggests that the two test statistics are measuring, at least partially, different properties of the data. Demographic events such as bottlenecks or population expansion resulted in a slightly higher correlation between lnRH and lnRV (Table 7).

In principle, both test statistics lnRV and lnRH have a type I error of 5%. Thus, in a microsatellite screen of 100 loci, five loci will be identified as putative targets of selection. We were interested to test, whether this high number of false positives could be reduced when lnRH and lnRV are considered jointly. Table 8 indicates the fraction of loci which were identified to be significant (α=0.05) for both lnRV and lnRH. Interestingly, the number of false positives was reduced about three-fold.

To further evaluate the combined lnRH-lnRV test statistic, we determined the power of this statistic to detect one selected locus out of 99 neutrally evolving ones. The identical simulation runs were used to determine the statistical power of each of the three test statistics. For an old selective sweep the power of the combined lnRH-lnRV test statistic was substantially lower than for each of the other two test statistics (Table 9). For strong and recent selective sweeps, however, the combined lnRH-lnRV test statistic had almost the identical power as lnRV. Despite that lnRH was the most powerful test statistic, the advantage of the combined lnRH-lnRV test is that the number of false positives is reduced by a factor of three.

Table 9. Power of lnRH, lnRV and the combined test statistic (CRHV)

	θ=3			θ=5			θ=10			θ=50		
	lnRH	**lnRV**	**CRHV**	**lnRH**	**lnRV**	**CRHV**	**lnRH**	**lnRV**	**CRHV**	**lnRH**	**lnRV**	**CRHV**
t_S=0.2, f_S=0.01	0.37	0.18	0.14	0.32	0.16	0.10	0.26	0.16	0.10	0.22	0.16	0.08
t_S=0.1, f_S=0.01	0.52	0.35	0.31	0.54	0.31	0.27	0.51	0.34	0.28	0.43	0.34	0.26
t_S=0.05, f_S=0.01	0.77	0.56	0.54	0.82	0.59	0.57	0.82	0.57	0.55	0.79	0.57	0.54
t_S=0.05, f_S=0.001	0.80	0.61	0.59	0.87	0.62	0.61	0.90	0.62	0.61	0.85	0.66	0.63

One locus was subjected to directional selection and 99 loci evolved neutrally. 1000 simulation runs were performed. See Tables 3, 4 for an explanation of t, f.

Discussion

We have introduced a new test statistic (lnRH) for the detection of genomic regions subjected to a recent selective sweep. Similar to the previously introduced lnRV test statistic, lnRH follows a Gaussian distribution for a broad range of parameters. The principle of both test statistics is that the occurrence of a recent selective sweep is expected to reduce variability at a microsatellite locus linked to the selected site, but other regions of the genome should not be affected. Hence, for both test statistics a selected locus is expected to have a lnRV or lnRH value that differs significantly from the remainder of the genome. Using the density function of the Gaussian distribution, it is possible to determine the probability that a locus differs from the remainder of the genome by chance. One very attractive property of both statistics is that the distribution of the test statistic captures the demographic history of the population. The largest caveat of both test statistics is that when a large number of loci are scored even for a neutrally evolving population, a large number of putative candidate loci will be identified.

While this problem could be accounted for by adjusting the experiment-wise error rate α, standard procedures such as a Bonferroni correction are extremely conservative resulting in a large type II error (false negatives).

The new lnRH test statistic offers the advantage of a smaller variance than lnRV, which significantly increases the power of the lnRH test statistic. Thus, a Bonferroni correction applied to lnRH results in a smaller type II error than for the lnRV test statistic. To further decrease the type I error, we jointly applied the lnRV and lnRH test to the same data set. Only if a microsatellite was identified as a significant outlier by both test statistics, was this locus considered to deviate from neutral expectations. Using this strategy, we observed approximately a three-fold reduction in rate of false positives.

For species with a fully sequenced genome, it is possible to verify a putative selective sweep by the analysis of flanking microsatellites and DNA sequencing.[6] For nonmodel organisms, however, this strategy is not feasible. In such cases the combined lnRH-lnRV test statistic could provide some additional confidence on microsatellite loci deviating from the remainder of the genome. Nevertheless, note that deviations from the strict stepwise mutation model affect both test statistics differently, precluding the routine use of the combined lnRH-lnRV test statistic.

Acknowledgments

We are grateful to Max Kauer and the rest of the CS lab for stimulating discussions. M. Kauer and G. Muir provided helpful comments on the manuscript. This work has been supported by Fonds zur Förderung der wissenschaftlichen Forschung (FWF) grants to CS and a infrastructure grant for genomics and transciptomics awarded to the Veterinärmedizinische Universität Wien by the bm:bwk.

References

1. Hudson RR, Bailey K, Skarecky D et al. Evidence for positive selection in the superoxide dismutase (Sod) region of Drosophila melanogaster. Genetics 1994; 136:1329-1340.
2. Ferea TL, Botstein D, Brown PO et al. Systematic changes in gene expression patterns following adaptive evolution in yeast. PNAS 1999; 96(17):9721-9726.
3. Toma DP, White KP, Hirsch J et al. Identification of genes involved in Drosophila melanogaster geotaxis, a complex behavioral trait. Nat Genet 2002; 31(4):349-353.
4. Gould SJ, Lewontin RC. The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. Proc Roy Soc Lond B 1979; 205:581-598.
5. Maynard Smith J, Haigh J. The hitch-hiking effect of a favorable gene. Genet Res 1974; 23:23-35.
6. Harr B, Kauer M, Schlötterer C. Hitchhiking mapping—a population based fine mapping strategy for adaptive mutations in D melanogaster. PNAS 2002; 99:12949-12954.
7. Kim Y, Stephan W. Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 2002; 160(2):765-777.

8. Schlötterer C. Towards a molecular characterization of adaptation in local populations. Curr Opin Genet Dev 2002; 12(6):683-7.
9. Kohn MH, Pelz HJ, Wayne RK. Natural selection mapping of the warfarin-resistance gene. PNAS 2000; 97(14):7911-7915.
10. Wootton JC, Feng X, Ferdig MT et al. Genetic diversity and chloroquine selective sweeps in Plasmodium falciparum. Nature 2002; 418(6895):320-323.
11. Vigouroux Y, McMullen M, Hittinger CT et al. Identifying genes of agronomic importance in maize by screening microsatellites for evidence of selection during domestication. PNAS 2002; 99(15):9650-9655.
12. Payseur BA, Cutter AD, Nachman MW. Searching for evidence of positive selection in the human genome using patterns of microsatellite variability. Mol Biol Evol 2002; 19(7):1143-1153.
13. Schlötterer C, Vogl C, Tautz D. Polymorphism and locus-specific effects on polymorphism at microsatellite loci in natural Drosophila melanogaster populations. Genetics 1997; 146:309-320.
14. Schlötterer C. A microsatellite-based multilocus screen for the identification of local selective sweeps. Genetics 2002; 160(2):753-763.
15. Harr B, Zangerl B, Brem G et al. Conservation of locus specific microsatellite variability across species: A comparison of two Drosophila sibling species D melanogaster and D simulans. MBE 1998; 15:176-184.
16. Di Rienzo A, Donnelly P, Toomajian C, et al. Heterogeneity of microsatellite mutations within and between loci and implications for human demographic histories. Genetics 1998; 148:1269-1284.
17. Hudson RR. Gene geneologies and the coalescent process. Oxf Surv Evol Biol 1990; 7:1-44.
18. Ohta T, Kimura M. A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet Res 1973; 22:201-204.
19. Di Rienzo A, Peterson AC, Garza JC et al. Mutational processes of simple-sequence repeat loci in human populations. PNAS 1994; 91:3166-3170.
20. Lynch M, Walsh B. Genetics and analysis of quantitative traits. Sunderland: Sinauer Associates, 1998.
21. Wiehe T. The effect of selective sweeps on the variance of the allele distribution of a linked multi-allele locus-hitchhiking of microsatellites. Theor Popul Biol 1998; 53:272-283.

CHAPTER 6

Detecting Hitchhiking from Patterns of DNA Polymorphism

Justin C. Fay and Chung-I Wu

The genetic basis of adaptive evolution has long escaped the grasp of evolutionary geneticists due to the difficulty of mapping an organism's phenotype to its genotype. However, adaptive substitutions may also be identified by their effects on linked neutral variation. This has made it possible to test whether an adaptive substitution has recently occurred in a particular gene and whether such substitutions are common within an organism's genome. Of critical importance is the power of tests that detect adaptive substitutions and our confidence in the evidence for such events.

Adaptive substitution can be detected by their effects on levels and patterns of DNA polymorphism. With few exceptions all tests compare some feature of observed polymorphism data with that expected under a Wright-Fisher neutral model. This model assumes mutations arise in a diploid population of size N with probability μ per generation, mating is random, there is no selection, there is no population structure, population size is constant, there are nonoverlapping generations, and the population is at mutation-drift equilibrium.[1] Although it is true that natural populations violate most of these assumptions, the neutral model is often sufficient to describe most features of polymorphism data obtained from natural populations. This is in part due to the fact that slight violations of these assumptions do not cause large deviations from the neutral expectation and in part because under neutrality nearly all features of polymorphism data are expected to be quite variable.

In this chapter we describe how various aspects of polymorphism data can be used to detect the effect of positive selection on linked neutral variation, or the hitchhiking effect. We also compare these methods, with respect to their power to detect hitchhiking and their sensitivity to violations of the Wright-Fisher model.

Reduction in Levels of Variation

The primary effect of positive selection on linked neutral variation is a reduction in heterozygosity (Fig. 1). In the absence of recombination, variation is steadily removed by hitchhiking or the spread of an advantageous allele through a population. Subsequent to hitchhiking variation is slowly regained by the drift of new mutations to detectable frequencies. When selection is strong the advantageous allele is fixed in approximately $\ln(2N)(2/s)$ generations, compared to a neutral allele which is expected to take $4N$ generations, where N is the effective population size and $1/2N$ is the initial frequency of the advantageous mutation.[2] Subsequent to a hitchhiking event most variation is regained within $4N$ generations.[3,4]

Selective Sweep, edited by Dmitry Nurminsky. ©2005 Eurekah.com and Kluwer Academic/Plenum Publishers.

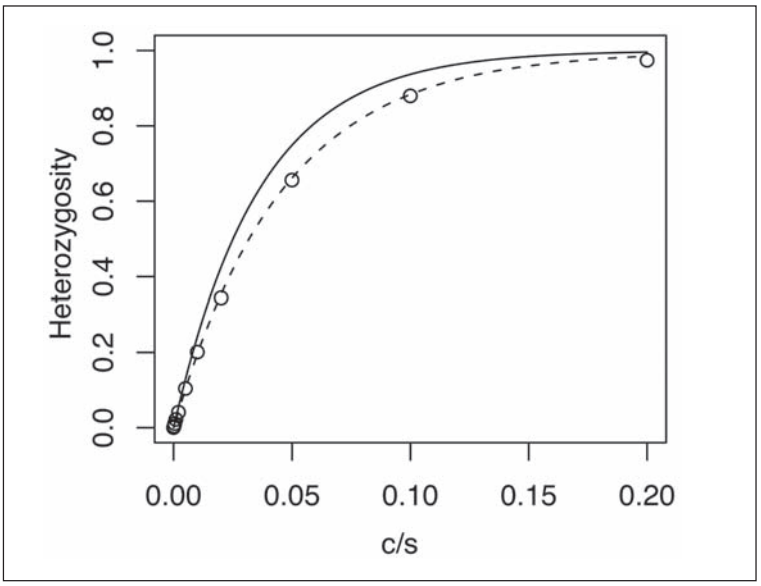

Figure 1. Heterozygosity as a function of c/s for the deterministic approximation of Maynard-Smith and Haigh,[36] eq. 8 $\approx 1-e^{2c/s}$ (solid line), the deterministic approximation of Stephan et al[45] eq. 17 (dashed line), and for 10^4 coalescence simulations (circles). Simulation parameters are $2N = 10^8$, $s = 10^{-3}$, $\varepsilon = 10^{-6}$, where ε is the initial frequency of the advantageous mutation.

In the presence of recombination, the reduction in heterozygosity is a function of the ratio of the rate of recombination to the selection coefficient, c/s, and the initial frequency of the advantageous mutation, assuming the spread of the advantageous mutation is deterministic.[5] This assumption is justified when the frequency of an advantageous mutation is greater than ε but less than $1-\varepsilon$, where ε is the frequency at which the probability the advantageous mutation is lost is nearly zero, i.e., $(1-2s)^{2N\varepsilon} \approx e^{-4Ns\varepsilon} \approx 0$, where 1, $1+s$ and $1+2s$ are the fitnesses of genotypes aa, Aa and AA, respectively.[6] Various approximations have been made to account for the hitchhiking dynamics below ε and above $1-\varepsilon$,[6-9] but if selection is strong, the stochastic phase of the hitchhiking event does not have much influence on the time to fixation.[7] However, it should be noted that recombination events that occur when the advantageous mutation is rare can have a large effect on the reduction in heterozygosity at a nearby locus. Thus, even a slight change in the time spent between $1/2N$ and ε is expected to magnify or reduce the effects of recombination on hitchhiking.[7]

A reduction in heterozygosity can be used as evidence for hitchhiking. The HKA test[10] detects a reduction in heterozygosity at one locus compared to a reference locus, and the test has been applied to many genes in *Drosophila melanogaster*.[11] Although the test accounts for different mutation rates at different loci within the genome, the results can be difficult to interpret since the significance of the test varies depending on which "neutral" locus is used as a reference. The HKA test is also sensitive to population subdivision, which increases the variance in heterozygosity across the genome,[12] and to purifying selection which is expected to reduce levels of variation as a function of the recombination rate and of the rate of deleterious mutations.[13] More compelling arguments for hitchhiking can be made by showing a local reduction in variation along a chromosome (as shown in Fig. 1). This has been done for the *Acp26Aa,*[14,15] *Sod*[16,17] and *Sdic*[18] genes in *D. melanogaster*. However, even under a neutral model, a local reduction in levels of variation may be observed due to the large evolutionary

variance in the time to the most recent common ancestor. The difficulty lies in determining how large a region and how great of a reduction in levels of variation cannot be explained by a neutral model. Kim and Stephan[19] have developed a maximum likelihood method to test for hitchhiking based on polymorphism sampled along a chromosome. The test is based on both a reduction in levels of variation and a skew in the frequency spectrum.

Skew in the Frequency Spectrum

The effect of hitchhiking on the frequency spectrum depends on the ratio of the recombination rate to the selection coefficient, the initial frequency of the advantageous mutation, and most importantly on the time since the start (or end) of the hitchhiking event. During the spread of an advantageous mutation, neutral mutations are swept to either low or high frequency depending on their original linkage relationship with the advantageous mutation. In the absence of recombination, a partial hitchhiking event (where the advantageous mutation does not reach fixation), can be detected by a single mutation or haplotype present at a much higher frequency than expected under a neutral model (see below). If there is no recombination and hitchhiking is complete, all variation is removed from a locus.

A skew in the frequency spectrum can also be produced as an indirect byproduct of removing all variation from a locus. Subsequent to hitchhiking, new mutations accumulate at low frequency in a population and it takes some time before they drift to intermediate or high frequencies. This skew in the frequency spectrum towards low frequency variation can be measured by Tajima's D statistic.[20] Tajima's D is the difference between two estimators of the population parameter θ divided by the standard deviation of the difference. Under the Wright-Fisher model the expectation of θ is equal to $4N\mu$, where N is effective population size and μ is the mutation rate. The two estimators are

$$\theta_\pi = \sum_{n=1}^{n-1} \frac{2S_i i\left(n-i\right)}{n\left(n-1\right)} \tag{1}$$

which is based on the average heterozygosity[21] and

$$\theta_W = \sum_{n=1}^{n-1} S_i \left(\sum_{n=1}^{n-1} \frac{1}{i} \right)^{-1} \tag{2}$$

which is based on the number of segregating sites divided by a constant, which depends on the the sample size n.[22] $\hat{\theta}_\pi$ is most sensitive to intermediate frequency variation, whereas $\hat{\theta}_w$ is most sensitive to rare (low or high frequency) variation. The reasoning is as follows: a single segregating sites at intermediate frequency adds 10×(20-10)/(20×19) = 0.26 to $\hat{\theta}_\pi$ whereas a low frequency variant adds much less: 1×(20-1)/380 = 0.05. In contrast, each segregating site contributes equally to $\hat{\theta}_w$. Since most variation in a population is found at low frequencies $\hat{\theta}_w$ is easily influenced by changes in the number of low frequency variants.

Under neutrality, the means of two estimators are expected to be equal to one another. Subsequent to a hitchhiking event that has removed all variation $\hat{\theta}_w$ is expected to be greater than $\hat{\theta}_\pi$ until new mutations reach intermediate frequency in a population. Simulation studies of hitchhiking events have shown that Tajima's D has quite a bit of power to detect a strong hitchhiking event $0.2N$ generations subsequent to the fixation of an advantageous mutation.[3] The advantage of this test is that no assumptions are made about how much variation is expected in a population. The disadvantage of this test, as well as all other tests that use polymorphism data, is that while recombination doesn't affect the mean it does affect the variance of the frequency spectrum and of test statistics based on the frequency spectrum. Recombination decreases the variance since it enables different mutations within a sample to have different

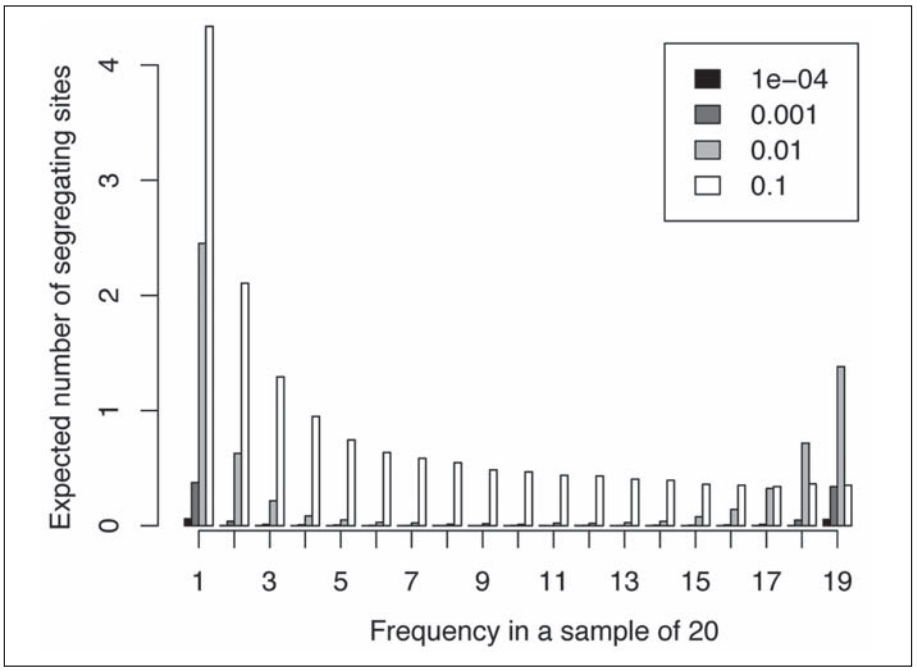

Figure 2. Expected frequency spectrum of sites in a sample of 20 subsequent to a hitchhiking event for different *c*/*s* values. Parameters are 10^4 coalescence simulations, $2N = 10^8$, $s = 10^{-3}$, $\theta = 5$, sample size is 20.

genealogies. While the rate of recombination can be either measured in the lab or estimated from polymorphism data, these estimates rely on a number of assumptions and often have large confidence intervals.[23] The practical solution that is most often taken is to conservatively assume no recombination for the purpose of generation of the cutoff values for a test statistic, or to use a conservative estimate of the recombination rate, typically the lower bound estimate.

A number of other tests, besides Tajima's D, have been developed to detect hitchhiking based on a skew in the frequency spectrum. Fu and Li's statistics D_{FL} and D^*_{FL}, test for a difference between $\hat{\theta}_\pi$ and θ estimated from the number of singletons (those mutations found only once in a sample). For D^*_{FL}, an outgroup is used to distinguish whether the derived mutation is found once or n-1 times in a sample of n. To provide a general framework for comparisons between the observed frequency spectrum and the neutral expectation, Fu derived an estimate of θ for every frequency class in a sample; $\theta_i = iS_i$.[24] Comparison of the frequency based tests showed that Tajima's D has the most power to detect a hitchhiking event in the absence of recombination.[25]

In the presence of recombination, hitchhiking produces a skew in the frequency spectrum quite different from that in the absence of recombination. In the presence of recombination a neutral variant will increase or decrease in frequency depending on whether it belongs to the same haplotype as the advantageous mutation or not. For a deterministic hitchhiking event, the expected final frequency of a neutral variant depends on the ratio of the rate of recombination to the selection coefficient and on the initial frequency of the advantageous mutation.[5] The end result is that subsequent to a strong hitchhiking event, neutral variation that has recombined into the advantageous haplotype is found at either high or low frequencies and thus forms a bipartite frequency spectrum (Fig. 2).[15] High and low frequency variation refer to

the frequency of the derived variant (or new mutation) which is distinguished from the ancestral variant using an outgroup. Subsequent to the hitchhiking event, high frequency variants are lost and new mutations at low frequency accumulate.[26-28]

The bipartite frequency spectrum produced in the presence of recombination can be detected by Tajima's D statistic,[15] or any other statistic that measures differences between rare and common variation. However, low frequency variation is easily influenced by changes in population size and by background selection (see below). On the other hand, an excess of high frequency as compared to common frequency variation cannot easily be produced by demographic scenarios (see below). $\hat{\theta}_H$ is a measure of high frequency variation and is based on the homozygosity of the derived variant.

$$\theta_H = \sum_{n=1}^{n-1} \frac{2S_i i^2}{n(n-1)} \tag{3}$$

The H test is the difference between $\hat{\theta}_\pi$ and $\hat{\theta}_H$, and is therefore a test for an excess of high frequency as compared to intermediate frequency mutations.[15] Because an outgroup must be used to distinguish between high and low frequency mutations, the probability of mis-inference must be incorporated into applications of the H test. The derived state can be mis-inferred if a reverse mutation occurs at a site. If all sites have the same mutation rate and thus the same probability of a reverse mutation, the probability of mis-inference can be estimated by $d/3$, where d is the rate of divergence corrected for multiple hits and 1/3 is the probability that a mutation is a reverse mutation, A to T, rather than A to G, when A and T are segregating.[15] Differences in the rate of transitions and transversions or other mutational biases can also be incorporated.[15]

Both Tajima's D and the H test have good power to detect hitchhiking in the presence of recombination (Fig. 3). In contrast to D the power of H drops rapidly after the hitchhiking event since high frequency variants as measured by $\hat{\theta}_H$ are readily lost due to drift.[26-28] Tajima's D retains power for much longer due to the influx of new low frequency variation during the recovery from a hitchhiking event (Fig. 3). Because variation is recovered first at low, then intermediate, and then high frequencies, a test for a lack of high frequency variation may retain the most power for the longest period of time subsequent to a hitchhiking event. The difference between $\hat{\theta}_H$ and $\hat{\theta}_W$, H_L, is a measure of high frequency compared to low frequency variation, and retains power for the longest period of time subsequent to hitchhiking (Fig. 4). This can be explained by $\hat{\theta}_H$ being the last of the three estimators of θ to reach equilibrium and $\hat{\theta}_W$ being the first to reach equilibrium.

Using the expected reduction in heterozygosity in combination with the expected skew in the frequency spectrum in the presence of recombination, Kim and Stephan[19] have implemented a maximum likelihood approach to simultaneously test for hitchhiking and to estimate both the location of the advantageous substitution and the strength of selection, given the recombination rate. Although this test appears more powerful than the tests based on different estimators of θ, it requires precise knowledge of the recombination rate and may be more sensitive to nonequilibrium conditions, since the null and the alternative hypotheses are more precisely specified. Yet, it should be noted that the robustness of all tests to violations of the assumptions of the Wright-Fisher model has not been well characterized (see below). In one of the first attempts to explicitly test selective versus demographic explanations, Galtier et al[29] have used a maximum likelihood approach to distinguish selection from a population bottleneck using data from *Drosophila* for which multiple loci have been surveyed for polymorphism. The logic behind the test is that a population bottleneck is expected to reduce levels of variation and skew the frequency spectrum across all loci, whereas a hitchhiking event is expected to be specific to only a fraction of loci.

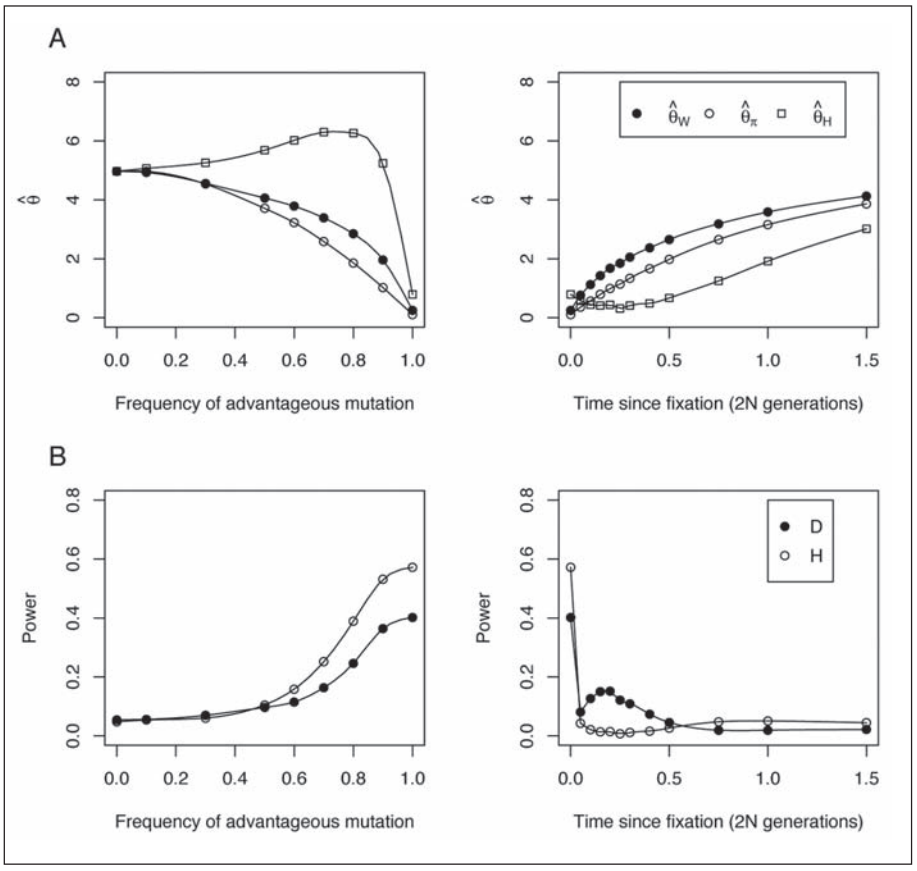

Figure 3. A) The expectation of different estimators of θ during and subsequent to hitchhiking. B) The power of the D and H statistics during and subsequent to hitchhiking. The simulation parameters are the same as in Figure 2 except *c*/*s* is fixed at 10^{-3}. For each simulated hitchhiking event with at least one segregating site D and H were compared to critical values generated from 10^4 neutral coalescence simulations with a fixed number of segregating sites equal to that observed in the hitchhiking simulation.

Linkage Disequilibrium

Hitchhiking is expected to produce linkage disequilibrium both in the presence and in the absence of recombination.[30] During the spread of an advantageous mutation through a population, a haplotype of very tightly linked neutral variants will increase in frequency until fixation. In some instances a second haplotype may remain segregating at appreciable frequencies (>1%) by recombining onto the advantageous chromosome during the hitchhiking event. Farther away from the site under selection, recombination events allow one or more different haplotypes to recombine onto the advantageous chromosome and thus escape extinction. As the distance to the site under selection increases, so does the number of alleles that escape complete hitchhiking (Fig. 3 of ref. 15). If the rate of recombination is low enough so that there is no recombination within the sequence surveyed, but high enough so that variation remains segregating subsequent to hitchhiking, then a strong haplotype pattern may form where all variation is divided among only a few haplotypes. In the extreme case where only two haplotypes remain segregating, all variation may be in complete linkage disequilibrium. A neutral

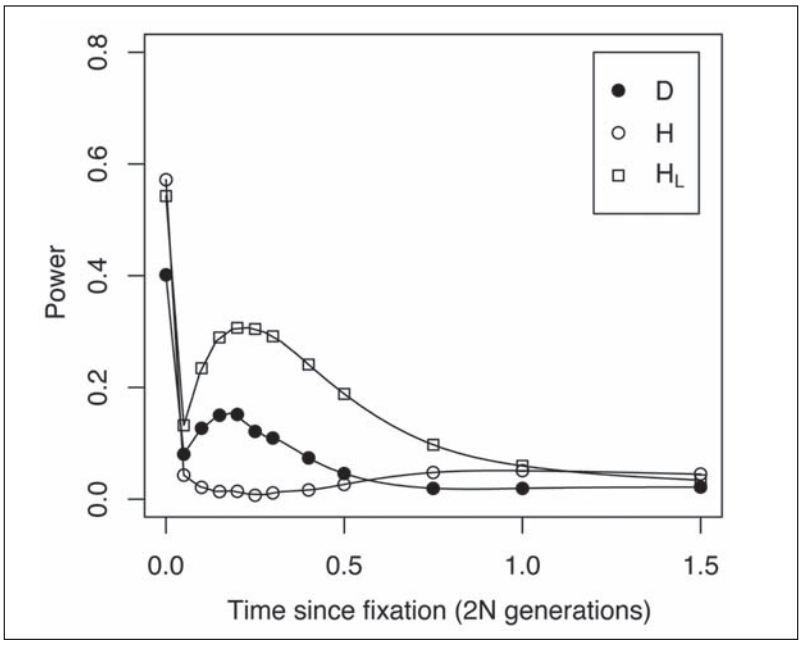

Figure 4. The power of D, H and H_L as a function of time since hitchhiking. H_L is the difference between $\hat{\theta}_W$ and $\hat{\theta}_H$. The simulation parameters are the same as those in Figure 3.

model may not be able to explain the presence of a single haplotype at intermediate or high frequency.[16,31] In addition to hitchhiking with recombination, a single haplotype could reach high frequency (but not fixation) due to balancing selection, the loss of positive selection during a hitchhiking event, or interference with advantageous or deleterious mutations in the population.[16,31] The degree to which hitchhiking produces linkage disequilibrium between two alleles can be measured by r (their correlation coefficient) and by D', the difference between the observed and expected (assuming independence) biallelic frequencies in a sample.[32]

$$r = \frac{f_{AB} - f_A f_B}{\sqrt{f_A f_B (1 - f_A)(1 - f_B)}} \tag{4}$$

$$D' = \frac{f_{AB} - f_A f_B}{\min\left[f_A f_B, (1 - f_A)(1 - f_B)\right]} \; for D' > 0 \tag{5}$$

$$D' = \frac{f_{AB} - f_A f_B}{\min\left[f_A (1 - f_B), (1 - f_A), f_B\right]} \; for D' < 0 \tag{6}$$

where f_A is the frequency of the major allele at the first locus, f_B is the frequency of the major allele at the second locus and f_{AB} is the frequency of the AB haplotype. Strong hitchhiking produces more linkage disequilibrium than expected in the absence of recombination, when

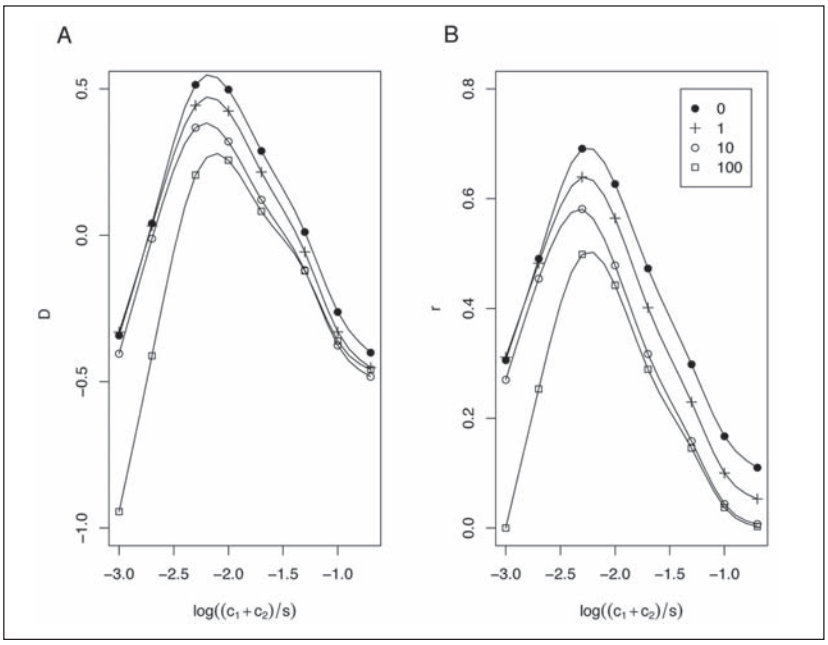

Figure 5. The average of D' (A) and r (B) as a function $(c_1+c_2)/s$, where c_1 is the rate of recombination between the selected locus and adjacent neutral locus and c_2 is the rate of recombination between the two neutral loci. $4Nc_2 = 0$ (solid circles), $4Nc_2 = 1$ (cross), $4Nc_2 = 10$ (open circles), $4Nc_2 = 100$ (squares), samples size is 50, $2N = 10^8$.

measured by r and D'.[19,26,28] This is true even when some recombination is allowed between the two neutral markers during hitchhiking (Fig. 5). However, previous work has shown that linkage disequilibrium decays rapidly subsequent to hitchhiking.[28] More work is necessary to distinguish linkage disequilibrium created by demographic effects or selection.

A number of haplotype tests have been developed to detect a high frequency haplotype or a lack of haplotype diversity that may occur during or subsequent to a hitchhiking event. Hudson et al[16] developed a test to determine the probability of observing a given number of segregating sites or fewer in a subset of sequences from a sample, and applied this to the *Sod* locus. The Fs test[25] is equal to $\ln(S/(1\text{-}S))$, where S is the probability of having no fewer than k alleles in a sample given $\hat{\theta}_\pi$.[33] Depaulis and Veuille[33] have proposed two tests for an excess of linkage disequilibrium (see also their chapter in this book). One is based on haplotype diversity, and another K, is based on the number of haplotypes, and both are conditioned on the number of segregating sites in a sample. K and Fs are only different in that they are conditioned on different estimators of θ.

Population Subdivision and Changes in Population Size

The effect of hitchhiking on linked neutral variation in a structured population, or in one that has recently changed its size, is not easily understood. However, in most cases the qualitative dynamics of hitchhiking are expected to be the same; variation is removed from a population producing a skew in the frequency spectrum and a linkage disequilibrium. Hitchhiking in a structured population is particularly difficult to describe since it depends on the number of subpopulations, the migration rates between subpopulations, and the effective size of these subpopulations. When the number of emigrants is less than one per generation, it has been

shown that hitchhiking causes population differentiation as a function of the strength of selection.[34] The effect of hitchhiking in a two-dimensional model of isolation by distance has also been studied.[8]

More important than understanding how hitchhiking is affected by population structure or changes in population size, is how the assumption of a constant panmictic population affects current methods of detecting hitchhiking. If demographic forces produce patterns that resemble hitchhiking, then the rate of erroneously detecting a hitchhiking event (i.e., rate of false positives) may be high. If demographic forces produce a pattern opposite to that of hitchhiking, then the power of detecting hitchhiking (rate of true positives) may be low. For all of the above mentioned tests, the rate of true and false positives is affected by both population subdivision and changes in population size. This results both from the effect of demography on the expectation of statistics such as Tajima's D, but also from the effect of demography on the variance in D. Selective forces are often distinguished from demographic forces by virtue of fact that the former is expected to be locus specific, while the latter is expected to affect the entire genome. However, if demography has a slight effect on the mean value of a test statistic or only affects the variance of a test statistic, it is likely to remain unnoticed as long as only a few loci across the genome are examined. Thus, it is important to know how changes in population size and population subdivision affect various tests used to detect hitchhiking.

A change in population size affects levels of variation, the frequency spectrum, and linkage disequilibrium. An increase in population size causes an increase in levels of low frequency variation and results in a negative Tajima's D value, whereas a decrease in population size causes a decrease in levels of low and high frequency variation and leads to positive Tajima's D.[35] The variance in Tajima's D has been shown to decrease in an expanding population[36] and is likely increased in a shrinking population. An increase in population size also causes a decrease in linkage disequilibrium as measured by the r statistic.[37]

Population structure affects patterns of variation in a much more complicated way. Tajima[35] studied a simple model of two demes with balanced migration. In the case where samples are drawn from both subpopulations, the heterozygosity increases faster than the number of segregating sites as the rate of migration decreases, thereby producing positive Tajima's D values. If samples are drawn from just one of the subpopulations, heterozygosity remains unchanged while the number of segregating sites decreases slightly with intermediate rates of migration, $4Nm \approx 1$, producing slightly negative Tajima's D values. In contrast, with unbalanced migration where the rate of migration from one population is 19 times greater than from the other, the number of segregating sites increases faster than heterozygosity as rates of migration decrease, when samples are drawn from both populations. Wakeley[12] found the variance in heterozygosity both within and between populations increases with the migration rate for a two subpopulation model with balanced migration. Population subdivision is also known to increases levels of linkage disequilibrium.[38]

Although few statistics have been tested for sensitivity to different population histories, there are obvious cases in which certain events in a population's history would mimic hitchhiking. For Tajima's D and Fu and Li's D_{FL} this would be a recent increase in population size, for the H test—the presence of a rare migrant from a distantly related population or species, and for the haplotype based tests—population subdivision or recent admixture. One case has been studied for the Tajima's D and for the H test. For a two-subpopulation model with balanced migration where 50 alleles are sampled from a single subpopulation, Tajima's D is significant in 6% of cases when $4Nm = 1$ and in 9% of cases when $4Nm = 0.5$, whereas the H test is significant in 14% and 19% of cases for 4Nm = 1 and 0.5, respectively.[28] However, under most circumstances the D and H tests would not be applied to a sample from a single isolated population. When sample are drawn from a mixture of subpopulations, the D and the H statistics are likely conservative because subdivision tends to produce an excess of intermediate frequency variation as compared to low frequency variation.

The simplest way of distinguishing demographic from selective effects is by surveying other unlinked loci in the genome. Any demographic perturbation is expected to affect all loci, whereas selection is expected to be specific to only a few loci. Subtle demographic effects, such as an increase in the variance of a statistic, are the most worrisome since they may remain unnoticed in a survey of a small number of genes but may still affect the rate of false positives of a test. Multiple independent lines of evidence, such as a regional reduction in levels of variation in combination with a skew in the frequency spectrum should be used to rule out a demographic explanation.

Distinguishing Background Selection and Hitchhiking in Regions of Low Recombination

One of the few genome wide patterns in polymorphism data that cannot be attributed to mutation and drift is the correlation between levels of variation and rates of recombination. This observation has now been made in numerous species, but it is still debated as to its cause.[39] The observation cannot be explained by different mutation rates, because rates of recombination are not correlated with divergence between species. However, there is accumulating evidence for heterogeneity in levels of divergence between two species, suggesting mutation rates may vary across the genome.[40] A question that has not been answered is the extent to which heterogeneity in levels of variation across the genome can be explained by mutational heterogeneities alone. The effect of regional differences in mutation rates across the genome must be accounted for in explaining low levels of variation in regions of low recombination.

Both background selection and recurrent hitchhiking can produce reduced levels of variation in regions of low recombination. With a sufficiently high rate of deleterious mutations per cM, background or purifying selection against deleterious mutations removes linked neutral variation, essentially reducing a population's effective size.[13] With a sufficiently high rate of adaptive substitutions driven by sufficiently strong selection, recurrent hitchhiking events may also maintain low levels of variation across an entire region of low recombination.[41] Tajima's D statistic is often used to distinguish between background selection and hitchhiking.[42] Simulation studies have shown that recurrent hitchhiking events in the presence of recombination produce an excess of low frequency variants and significantly negative D values.[41] In contrast, simulation studies have shown that background selection produces little or no skew in the frequency spectrum if Ns is sufficiently large, where s is the strength of selection against deleterious mutations.[13,33,43] When background selection does affect the frequency spectrum, Fu and Li's D has the most power to detect it.[25] Numerous polymorphism surveys were conducted in regions of low recombination with the aim of distinguishing background selection from hitchhiking by means of a skew in the frequency spectrum as measured by Tajima's D.[42-49] However, in many of these cases there was so little variation found that there was no power to detect a significant skew in the frequency spectrum.

If selection is so weak that deleterious mutations reach detectable frequencies (>1%) in a population, these mutations and neutral mutations linked to them are expected to produce an excess of low frequency variation as compared to common variation. Studies of allozyme variation in humans and fruit flies indicate that a large proportion of low frequency amino acid variants are slightly deleterious and that they reach detectable frequencies in a population.[50] By comparing the distribution of amino acid variation to synonymous variation, demographic explanations were ruled out and many of these deleterious mutations were shown to reach frequencies of 1-10% for both humans[51] and *D. melanogaster.*[52] Forward simulations of purifying selection show that mutations with $2Ns$ values as small as 6 can reduce levels of variation and produce negative D values in the absence of recombination.[53] The same effect is found when deleterious mutations are gamma distributed and there is no recombination.[54] Thus, at

least in the absence of recombination, background selection may produce negative D values as long as a sufficient number of slightly deleterious mutations are present.

The H test can be used to distinguish hitchhiking and background selection in regions of low recombination. The H statistic should not be affected by background selection, which only skews the frequency spectrum at low frequencies. In fact, in the presence of background selection hitchhiking may produce a larger excess of high frequency variants as compared to intermediate frequency variants than in the absence of background selection. The greater number of high frequency variants is the result of the excess of low frequency variants present prior to hitchhiking. It is these low frequency variants that are swept to high frequencies during hitchhiking. Thus, under the extreme example where only low frequency variants are present in a population, hitchhiking may produce only high frequency variants since all low frequency variants are either swept to high frequency or to frequencies too low to be detected. There are a number of regions where this has been observed. For example, the *y-ac* region located at the tip of the X chromosome of *D. melanogaster* shows three high frequency RFLP variants.[15] Similarly, five olfactory receptor pseudogenes in a 450 kb region of the human genome contain predominantly high frequency variants.[55]

To distinguish background selection from hitchhiking the H test must have reasonable power to detect recurrent hitchhiking events. Recurrent hitchhiking is different from a single hitchhiking event since at the start of each hitchhiking event the population is not at equilibrium. In most instances the population is likely recovering from the last hitchhiking event and so should have an excess of low frequency variants. The next hitchhiking event is expected to sweep low frequency variation to high or lower frequencies. Although coalescence simulations show that the H test has little power to detect recurrent hitchhiking events, this has been studied only for very strong selection and infrequent hitchhiking events, thus imposing a limitation on the approach.[28] Under these conditions, the power of detecting hitchhiking using the H test drops quickly subsequent to the fixation of the advantageous mutation. However, as the frequency of hitchhiking events increases, the neutral frequency spectrum may approach a U shaped distribution, which is the expected frequency distribution for mutations under positive selection.[1]

Finally, background selection and hitchhiking may be distinguished in a subdivided population if hitchhiking occurs exclusively or predominantly in only some of the subpopulations.[56] Background selection is expected to have similar effects in all subpopulations, whereas hitchhiking may be subpopulation-specific. For example, the *vermilion* locus was shown to have significantly reduced levels of variation in two out of four subpopulations of *D. ananassae*.[56]

Conclusions and Future Directions

Significant advances have been made in detection of positive selection using DNA polymorphism data. While a slew of new test statistics have been developed and shown to have power to detect hitchhiking, it is a standard practice to assume no recombination in a randomly mating Wright-Fisher population for determining the cutoff values for these tests. As genomic surveys of polymorphism become available, reliable estimates of the recombination rate and populations' demographic history can be made,[36,57] thus improving the use of existing tests and perhaps leading to the development of new ones. In the meantime, convincing evidence for hitchhiking must include multiple lines of evidence, such as a local reduction in levels of variation and a local skew in the frequency spectrum.

Genomic surveys of polymorphism will provide some indication of the number and location of loci in the genome that have recently experienced a hitchhiking event, and will clarify the relative contributions of background selection and hitchhiking to the reduction in levels of variation in regions of low recombination. Examination of high frequency variation will be particularly helpful here, since low frequency variation is similarly influenced by both background selection and hitchhiking.

References

1. Ewens WJ. Mathematical population genetics. Springer-Verlag: 1979.
2. Kimura M, Ohta T. The average number of generations until extinction of an individual mutant gene in a finite population. Genetics 1969; 63(3):701-709.
3. Simonsen KL, Churchill GA, Aquadro CF. Properties of statistical tests of neutrality for DNA polymorphism data. Genetics 1995; 141(1):413-429.
4. Wiehe THE, Stephan W. Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from Drosophila melanogaster. Mol Biol Evol 1993; 10(4):842-854.
5. Maynard-Smith J, Haigh J. The hitch-hiking effect of a favourable gene. Genet Res 1974; 23(1):23-35.
6. Kaplan NL, Hudson RR, Langley CH. The "hitchhiking effect" revisited. Genetics 1989; 123(4):887-899.
7. Barton NH. The effect of hitch-hiking on neutral genealogies. Genet Res 1998; 72:123-133.
8. Barton NH. Genetic hitchhiking. Philos Trans R Soc Lond B Biol Sci 2000; 355(1403):1553-1562.
9. Stephan W, Wiehe THE, Lenz MW. The effect of stongly selected substitutions on neutral polymorphism: Analytical results based on diffusion theory. Theor Popul Biol 1992; 41:237-254.
10. Hudson RR, Kreitman M, Aguade M. A test of neutral molecular evolution based on nucleotide data. Genetics 1987; 116(1):153-159.
11. Moriyama EN, Powell JR. Intraspecific nuclear DNA variation in Drosophila. Mol Biol Evol 1996; 13(1):261-277.
12. Wakeley J. The variance of pairwise nucleotide differences in two populations with migration. Theor Popul Biol 1996; 49(1):39-57.
13. Charlesworth B, Morgan MT, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. Genetics 1993; 134(4):1289-1303.
14. Aguade M, Miyashita N, Langley CH. Polymorphism and divergence in the Mst26A male accessory gland gene region in Drosophila. Genetics 1992; 132(3):755-770.
15. Fay JC, Wu C-I. Hitchhiking under positive Darwinian selection. Genetics 2000; 155(3):1405-1413.
16. Hudson RR, Bailey K, Skarecky D et al. Evidence for positive selection in the superoxide dismutase (Sod) region of Drosophila melanogaster. Genetics 1994; 136(4):1329-1340.
17. Hudson RR, Saez AG, Ayala FJ. DNA variation at the Sod locus of Drosophila melanogaster: An unfolding story of natural selection. Proc Natl Acad Sci USA 1997; 94(15):7725-7729.
18. Nurminsky D, Aguiar DD, Bustamante CD et al. Chromosomal effects of rapid gene evolution in Drosophila melanogaster. Science 2001; 291(5501):128-130.
19. Kim Y, Stephan W. Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 2002; 160(2):765-777.
20. Tajima F. The effect of change in population size on DNA polymorphism. Genetics 1989; 123(3):597-601.
21. Tajima F. Evolutionary relationship of DNA sequences in finite populations. Genetics 1983; 105(2):437-460.
22. Watterson GA. On the number of segregating sites in genetical models without recombination. Theor Popul Biol 1975; 7(2):256-276.
23. Andolfatto P, Przeworski M. A genome-wide departure from the standard neutral model in natural populations of Drosophila. Genetics 2000; 156(1):257-268.
24. Fu YX. Statistical properties of segregating sites. Theor Popul Biol 1995; 48(2):172-197.
25. Fu YX. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics 1997; 147(2):915-925.
26. Fay JC. Detecting natural selection from patterns of DNA polymorphism and divergence. PhD thesis, University of Chicago; 2001.
27. Kim Y, Stephan W. Joint effects of genetic hitchhiking and background selection on neutral variation. Genetics 2000; 155(3):1415-1427.
28. Przeworski M. The signature of positive selection at randomly chosen Loci. Genetics 2002; 160(3):1179-1189.
29. Galtier N, Depaulis F, Barton NH. Detecting bottlenecks and selective sweeps from DNA sequence polymorphism. Genetics 2000; 155(2):981-987.
30. Thomson G. The effect of a selected locus on linked neutral loci. Genetics 1977; 85(4):753-788.

31. Kirby DA, Stephan W. Haplotype test reveals departure from neutrality in a segment of the white gene of Drosophila melanogaster. Genetics 1995; 141(4):1483-1490.
32. Lewontin RC. The interaction of selection and linkage. I. General considerations heterotic models. Genetics 1964; 49:49-67.
33. Depaulis F, Veuille M. Neutrality tests based on the distribution of haplotypes under an infinite-site model. Mol Biol Evol 1998; 15(12):1788-1790.
34. Slatkin M, Wiehe T. Genetic hitch-hiking in a subdivided population. Genet Res 1998; 71(2):155-160.
35. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 1989; 123(3):585-595.
36. Pluzhnikov A, DiRienzo A, Hudson RR. Inferences about human demography based on multilocus analyses of noncoding sequences. Genetics 2002; 161(3):1209-1218.
37. Pritchard JK, Przeworski M. Linkage disequilibrium in humans: Models and data. Am J Hum Genet 2001; 69(1):1-14.
38. Wall JD. Detecting ancient admixture in humans using sequence polymorphism data. Genetics 2000; 154(3):1271-1279.
39. Andolfatto P. Adaptive hitchhiking effects on genome variability. Curr Opin Genet Dev 2001; 11(6):635-641.
40. Williams EJ, Hurst LD. Is the synonymous substitution rate in mammals gene-specific? Mol Biol Evol 2002; 19(8):1395-1398.
41. Braverman JM, Hudson RR, Kaplan NL et al. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics 1995; 140(2):783-796.
42. Andolfatto P, Przeworski M. Regions of lower crossing over harbor more rare variants in African populations of Drosophila melanogaster. Genetics 2001; 158(2):657-665.
43. Charlesworth D, Charlesworth B, Morgan MT. The pattern of neutral molecular variation under the background selection model. Genetics 1995; 141(4):1619-1632.
44. Begun DJ, Aquadro CF. Evolution at the tip and base of the X chromosome in an African population of Drosophila melanogaster. Mol Biol Evol 1995; 12(3):382-390.
45. Berry AJ, Ajioka JW, Kreitman M. Lack of polymorphism on the Drosophila fourth chromosome resulting from selection. Genetics 1991; 129(4):1111-1117.
46. Hamblin MT, Aquadro CF. High nucleotide sequence variation in a region of low recombination in Drosophila simulans is consistent with the background selection model. Mol Biol Evol 1996; 13(8):1133-1140.
47. Jensen MA, Charlesworth B, Kreitman M. Patterns of genetic variation at a chromosome 4 locus of Drosophila melanogaster and D. simulans. Genetics 2002; 160(2):493-507.
48. Langley CH, Lazzaro BP, Phillips W et al. Linkage disequilibria and the site frequency spectra in the su(s) and su(w(a)) regions of the Drosophila melanogaster X chromosome. Genetics 2000; 156(4):1837-1852.
49. Wayne ML, Kreitman M. Reduced variation at concertina, a heterochromatic locus in Drosophila. Genet Res 1996; 68(2):101-108.
50. Ohta T. Statistical analyses of Drosophila and human protein polymorphism. Proc Natl Acad Sci USA 1975; 72:3194-3196.
51. Fay JC, Wyckoff GJ, Wu C-I. Positive and negative selection on the human genome. Genetics 2001; 158(3):1227-1234.
52. Fay JC, Wyckoff GJ, Wu C-I. Testing the neutral theory of molecular evolution with genomic data from Drosophila. Nature 2001; 415(6875):1024-1026.
53. Gordo I, Navarro A, Charlesworth B. Muller's Ratchet and the Pattern of Variation at a Neutral Locus. Genetics 2002; 161(2):835-848.
54. Williamson S, Orive ME. The genealogy of a sequence subject to purifying selection at multiple sites. Mol Biol Evol 2002; 19(8):1376-1384.
55. Gilad Y, Segre D, Skorecki K et al. Dichotomy of single-nucleotide polymorphism haplotypes in olfactory receptor genes and pseudogenes. Nat Genet 2000; 26(2):221-224.
56. Stephan W, Xing L, Kirby DA et al. A test of the background selection hypothesis based on nucleotide data from Drosophila ananassae. Proc Natl Acad Sci USA 1998; 95(10):5649-5654.
57. Frisse L, Hudson RR, Bartoszewicz A et al. Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. Am J Hum Genet 2001; 69(4):831-843.

CHAPTER 7

Periodic Selection and Ecological Diversity in Bacteria

Frederick M. Cohan

Abstract

Biodiversity in the bacterial world is strongly influenced by "periodic selection," in which natural selection recurrently purges diversity within a bacterial population. Owing to the extreme rarity of recombination in bacteria, selection favoring an adaptive mutation eliminates nearly all the diversity within an ecotype (defined as the set of strains using about the same ecological niche, so that an adaptive mutant or recombinant out-competes to extinction strains from the same ecotype). Diversity within an ecotype is only transient, awaiting its demise with the next periodic selection event. Ecological diversity in bacteria is governed by three kinds of mutations (or recombination events). Niche-invasion mutations found a new ecotype, such that the new genotype and its descendants escape the diversity-purging effect of periodic selection from their former ecotype. Periodic selection mutations then make the different ecotypes more distinct by purging the diversity within but not between ecotypes. Lastly, speciation-quashing mutations may occur, which can extinguish another ecotype even after it has had several private, periodic selection events. For example, an ecotype that shares all its resources with another ecotype, albeit in different proportions, may be extinguished by an extraordinarily fit adaptive mutation from the other ecotype.

Sequence clusters, as determined by a variety of criteria, are expected to correspond to ecotypes. Sequence-based approaches suggest that a typical named species contains many ecotypes.

That periodic selection occurs in nature is evidenced by the modest levels of sequence diversity observed within bacterial species, levels that are too low to be explained by genetic drift. Also, a special kind of periodic selection event, driven by "adapt globally, act locally" mutations, is inferred when strains fall into discrete sequence clusters over most of their genomes, but are aberrantly homogeneous in a small chromosomal region. Beyond establishing a history of periodic selection, this pattern can help corroborate that a set of sequence clusters correspond to ecotypes.

Introduction

One half-century ago, a simple experiment changed the way we think about the power of natural selection in bacterial populations. The classic experiment of Atwood, Schneider, and Ryan[1] demonstrated the phenomenon of periodic selection, whereby diversity within a bacterial population is purged recurrently by natural selection. The principle is that in an entirely

Selective Sweep, edited by Dmitry Nurminsky. ©2005 Eurekah.com and Kluwer Academic/Plenum Publishers.

asexual population, each adaptive mutation precipitates a round of natural selection which, if successful, fixes not only the adaptive mutation but also the entire genome of the mutant cell. In the absence of recombination, the adaptive mutation is unable to enter into any other genetic background, and so selection favoring the adaptive mutation drags the entire genome associated with it to fixation.

A recent reenactment of this experiment, supported by data from modern molecular biology, illustrates the diversity-purging power of periodic selection.[2] Descendants of a single *Escherichia coli* cell were cultured without benefit of recombination, and were allowed to evolve. Diversity within the population was monitored over time by assaying the frequency of spontaneous mutants resistant to the bacteriophage T5. The frequency of resistant cells started at zero, gradually increased due to mutation for fifty or more generations, then abruptly dropped back to zero, and this pattern was repeated several times. As in the original periodic selection paper, the crashes in frequency of the marker were interpreted as the result of periodic selection. The model is that adaptive mutations occur within the majority population of cells (in this case, marked by T5 sensitivity), and that the adaptive mutant and its clonal descendants drive to extinction all other lineages in the population (T5-resistant and T5-sensitive alike). Thus, the rise of the adaptive mutant genotype is manifested in these experiments by the disappearance of the minority marker. In Notley-McRobb and Ferenci's recent paper, the interpretation of periodic selection was supported by coincident sequence changes at *mgc* and *mgl*, loci known to play a major role in adaptation to laboratory culture.

Periodic selection clearly has the potential to sweep the diversity within a strictly asexual population, descended from a single clone, in laboratory culture. However, these laboratory experiments do not necessarily predict the effect of selection in natural populations of bacteria. While the bacteria in these periodic selection experiments were engineered to be strictly asexual, bacteria in nature undergo recombination, albeit at an extremely low rate.[3-5] Thus, the diversity-purging effect of periodic selection may be diminished in natural populations. Also, while a periodic selection event can purge the diversity among the clonal descendants within a culture flask, it is not clear how broad a population would be purged of diversity in nature. For example, would all of *E. coli* be purged of diversity by one periodic selection event in nature?

Here I will describe the effects of periodic selection for natural populations of bacteria, and I will show how periodic selection plays a central role in the origin of ecological diversity in the bacterial world. I will demonstrate that recombination is typically too rare in bacteria to diffuse the diversity-purging effect of periodic selection. I will demonstrate that within a typical named species, there appear to be dozens of "ecotypes"— ecologically distinct populations that have their own private periodic selection events, so that each ecotype escapes the periodic selection events of all other ecotypes. I will show that the genetic changes allowing a genotype to escape periodic selection from its previous ecotype form the basis of bacterial speciation. Finally, I will show that periodic selection has actually occurred in natural populations of bacteria, and I will introduce a method for using genomic data to detect past periodic selection events.

The Nature of Recombination in Bacteria

The microcosms of periodic selection experiments were designed to be devoid of sex, but so far as we know, sexual recombination has been a part of every bacterial species' history.[5,6] That a bacterial species has engaged in recombination in its past may be demonstrated through a diversity of sequence-based tests.[7] For example, in the homoplasy test of Maynard Smith and Smith,[8] recombination is implicated when an improbably high number of nucleotide substitutions have occurred twice or more in different parts of the phylogeny. This and similar methods have demonstrated the existence of recombination in all bacterial species investigated.

The rate at which recombination occurs in nature has been estimated through several sequence-based approaches, including the extent to which different genes or gene segments yield different phylogenetic relationships among strains; congruence of phylogenies based on different gene segments indicates rare recombination. The recombination rates may be estimated separately for recombination within populations[9,10] and between populations.[4,11-13] Typically, the recombination rates range from nearly an order of magnitude less than mutation in some of the most clonal of bacteria (e.g., *Staphylococcus aureus*),[14] to about half an order of magnitude greater than mutation in some of the most frequently recombining bacteria (e.g., *Neisseria meningitidis*),[11] with the exception so far of *Helicobacter pylori*, which recombines at a much higher (but not yet determined) rate.[15] For example, a 450 bp segment of *N. meningitidis* undergoes recombination at a rate of 1.2×10^{-6} per individual per generation, which is 3.6 times the mutation rate.[11,16]

While recombination in bacteria is rare, it is also promiscuous; bacteria are not fastidious about their choice of sexual partners. Homologous recombination occurs even between organisms that are 25% divergent in their DNA sequences.[17,18]

Also, bacteria can acquire novel genes and whole operons through heterologous recombination, sometimes from extremely divergent species. Lawrence and Ochman[19] have developed a method for identifying genes acquired from extremely distant sources: genes with a highly aberrant GC content are interpreted as foreign genes. Based on this principle, typically 5-15% of the genes in a bacterial genome appear to have been acquired from extremely distant relatives.[6]

Finally, recombination in bacteria is unidirectional, from a donor cell to a recipient cell, and usually only a small fraction of the genome is transferred.[5]

The Effect of Rare Recombination on Diversity within a Population

Consider next whether the rare recombination typical of bacteria should soften the diversity-purging effect of natural selection. Recombination can preserve the genetic diversity in two ways. First, if the adaptive mutation recombines into another genetic background, then the entire genome of the recipient is saved from extinction (Fig. 1, example A). Alternatively, if a segment from a strain lacking the adaptive mutation should recombine into a strain with the adaptive mutation, then that segment (only) will be saved from extinction (Fig. 1, example B).

I have investigated the relationship between recombination rate and the purging of diversity using a Monte Carlo method derived from a coalescence algorithm of Braverman et al.[20] In a completely asexual population, each cell after periodic selection will contain only DNA derived from the genome of the original adaptive mutation; with recombination, DNA derived from other cells existing at the time of the adaptive mutation can contribute to the genomes of cells surviving periodic selection. Figure 2 shows the diversity-purging effect of periodic selection over a range of recombination rates occurring in bacteria. When the intensity of periodic selection is strong (i.e., fitness advantage for the adaptive mutation is $s = 0.1$), each bout of periodic selection purges nearly all diversity within an ecotype. Over recombination rates observed in nature (from 0.1 to 3.6 times the mutation rate), periodic selection purges all but 0.0007% to 0.2% of the sequence diversity. Over more modest selection intensity (i.e., $s = 0.01$), periodic selection purges all but 0.02% to 2% of sequence diversity over naturally occurring recombination rates. It appears that recombination is ineffective at softening the diversity-purging effect of periodic selection in nature.

The Origins of Permanent Divergence

Each periodic selection event reduces the diversity within a population to little more than the clonal descendants of the original adaptive mutant. Therefore, the diversity accumulated

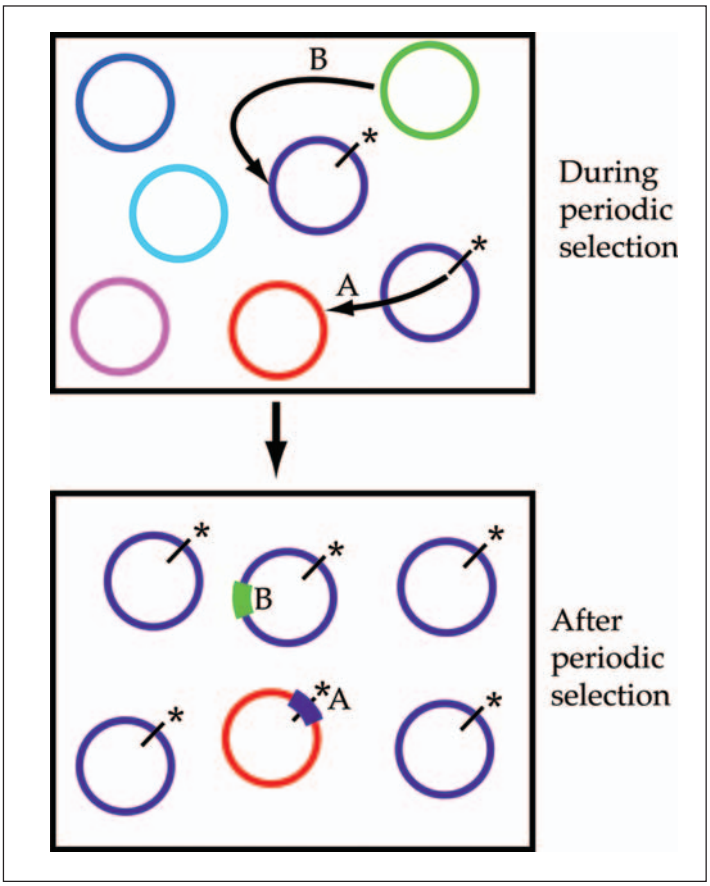

Figure 1. How recombination can potentially soften the diversity-purging effect of periodic selection. The adaptive mutation (indicated by an asterisk) originally appears in a cell with violet genetic background. Without recombination, only the genetic background of the original mutant would be represented in the population after periodic selection. A) If the adaptive mutation is transferred into another genetic background (shown in red), that background will survive the periodic selection (as seen in the lower panel). B) If a gene segment other than the locus of the adaptive mutation is transferred into a clonal descendant of the adaptive mutant, that segment (only) is saved from extinction. The recombination event saves a small segment of the red genetic background. The simulation on which (Fig. 2) is based takes into account these two kinds of recombination events, as well as recombination between strains bearing the adaptive mutation, and between strains without the adaptive mutation.

within a bacterial population is only transient, awaiting its demise with the next periodic selection event. What, then, is the evolutionary origin of permanent diversity in bacteria?

I have previously defined a bacterial "ecotype" with respect to the fate of an adaptive mutant: an ecotype is a set of strains using the same or similar ecological niches, such that an adaptive mutant from within the ecotype out-competes to extinction all other strains from the same ecotype; an adaptive mutant cannot, however, drive to extinction strains from other ecotypes (Fig. 3B).[18,21,22] Thus, an ecotype is the set of strains whose diversity is purged through periodic selection favoring each adaptive mutant. Periodic selection is a powerful force of cohesion within a bacterial ecotype, in that it recurrently resets the genetic diversity to near zero.

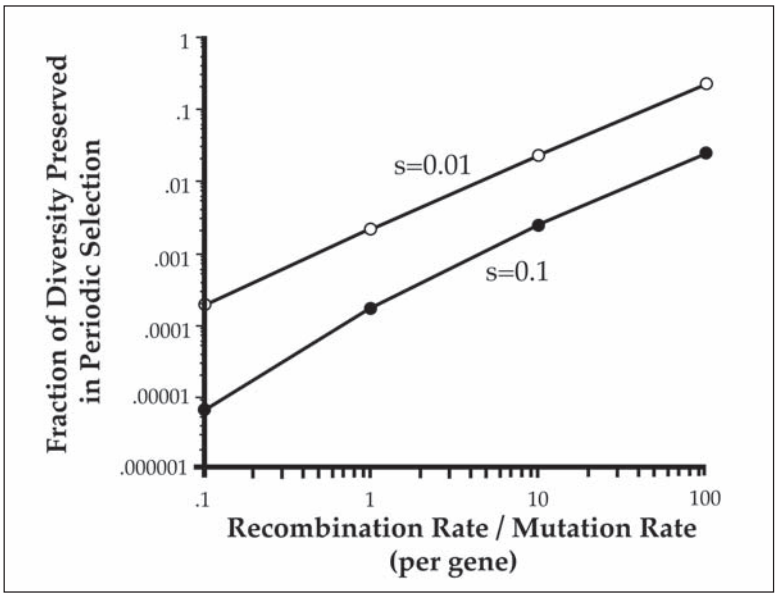

Figure 2. The relationship between recombination rate and the diversity-purging effect of periodic selection, over different intensities of selection (*s*) favoring the adaptive mutation. The ratios of recombination rate to mutation rate seen in the figure reflect the range of ratios observed in nature, with the exception of *Helicobacter pylori*, which recombines at a high but not yet determined rate. The ordinate is based on the mean fraction of a cell's genome at the end of periodic selection that is not descended from the genome of the original mutant. These results are based on a Monte Carlo simulation of coalescence of strains sampled from an ecotype at the end of periodic selection (based on Braverman et al[20]). Each point is based on 10,000 replicate runs, and standard error bars are too small to be visible.

At the point that two ecologically distinct populations undergo their own private periodic selection events, they have reached a milestone toward forming new species. Such populations are now irreversibly separate, since periodic selection cannot prevent further divergence (by definition), and as has previously been shown, neither can recombination.[21]

Even if recombination between ecotypes were to occur at the same rate as recombination within them, natural selection against rare inter-ecotype recombinants could easily limit the frequency of recombinant genotypes to negligible levels.[21] Therefore, evolution of sexual isolation is not a necessary step toward the evolution of permanent divergence in the bacterial world. The key milestone toward bacterial speciation is instead the genetic change that places a mutant (or recombinant) cell and its descendants outside the domain of periodic selection of other ecotypes.

Bacterial ecotypes, as defined by the domains of periodic selection, share the fundamental properties of species.[18,22] Ecotypes are each subject to an intense force of cohesion, periodic selection, which recurrently purges diversity within an ecotype (a species attribute emphasized by the Cohesion Species Concept of Templeton).[23] Once different ecotypes have diverged to the point of escaping one another's periodic selection events, there is no force that can prevent their divergence. (The irreversibility of divergence is emphasized by the Evolutionary Species Concept of Simpson[24] and Wiley).[25] As we shall see, ecotypes form distinct sequence clusters, owing to periodic selection purging sequence diversity within but not between ecotypes.[26] (The phenotypic and molecular separateness of species is emphasized by the Phenotypic Species Concept of Sokal and Crovello[27] and the Modern Synthesis Species Concept of Mallet.[21,28]) Finally, bacterial ecotypes are ecologically distinct, as emphasized by the Ecological

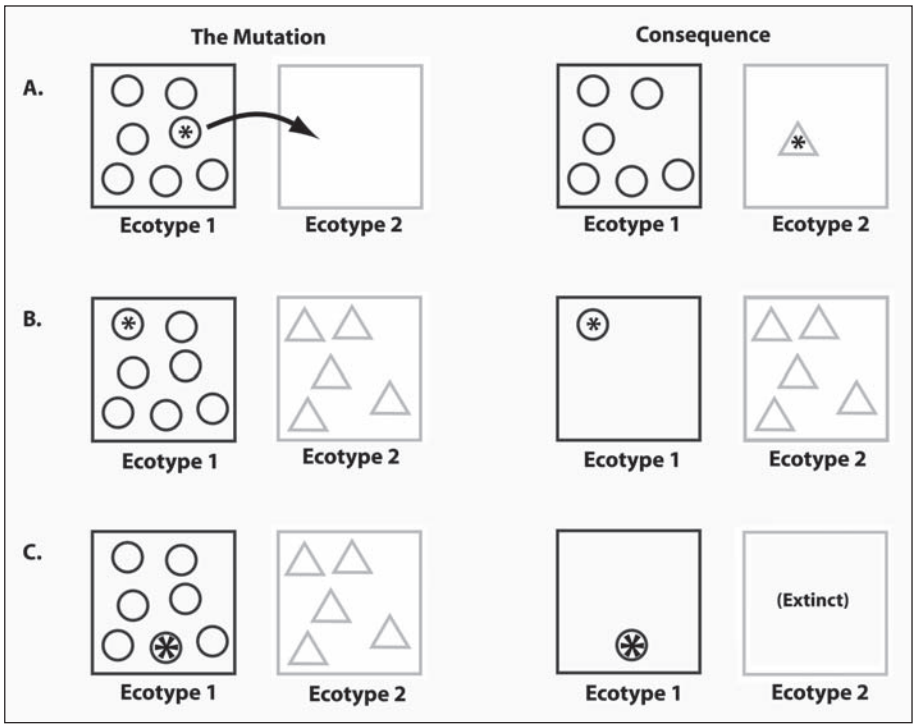

Figure 3. Three classes of mutation and recombination events that determine ecotype diversity in bacteria. The circles represent distinct genotypes, and the asterisks represent adaptive mutations. A) Niche-invasion mutations. Here a mutation changes the ecological niche of the cell, such that it can now escape periodic selection events in its former ecotype. This founds a new ecotype. B) Periodic-selection mutations. These improve the fitness of an individual such that the mutant and its descendants out-compete all other cells within the ecotype; periodic selection events precipitated by these mutations generally do not affect the diversity within other ecotypes, owing to the differences in ecological niche. Periodic selection enhances the distinctness of ecotypes by purging the divergence within but not between ecotypes. C) Speciation-quashing mutations. Even if two ecotypes have sustained a history of separate periodic selection events, an extraordinarily adaptive genotype may out-compete to extinction another ecotype. Competitive extinction of another ecotype (Ecotype 2) is possible only if all of Ecotype 2's resources are also used by Ecotype 1.

Species Concept of van Valen.[29] Bacterial ecotypes are therefore evolutionary lineages that are irreversibly separate, each with its own evolutionary tendencies and historical fate. A species in the bacterial world may be understood as an evolutionary lineage bound together by ecotype-specific periodic selection.[22]

Effects of Periodic Selection beyond the Boundaries of the Ecotype

The ecological divergence between ecotypes allows them to coexist and to survive each other's periodic selection events. Nevertheless, if newly divergent ecotypes compete for at least some resources, they may feel the effects of periodic selection from outside the ecotype. Suppose, for example, that a parental ecotype and a nascent ecotype use the same two sugars, but the parental ecotype takes up one sugar preferentially, while the reverse is true for the nascent ecotype. Modestly adaptive mutations that increase overall efficiency in the parental ecotype will fail to extinguish the nascent ecotype, but they can decrease its population density

significantly. Nevertheless, the genetic diversity of the nascent ecotype is not diminished (except minimally by increasing genetic drift). Provided that adaptive mutations are modest, these populations can coexist indefinitely, even as each adaptive mutation negatively impacts the population density of the other ecotype.

Even after newly divergent ecotypes have each undergone several rounds of their own, private periodic selection events, they may still be vulnerable to extinction caused by the other ecotype's periodic selection. This can be the case when ecotypes use entirely the same set of resources, but in different proportions. An extraordinarily fit adaptive mutant from the parental ecotype might out-compete all strains from the nascent ecotype (as well as all the other strains from its own ecotype) (Fig. 3C). In this case, the founding of the new ecotype would be quashed by a periodic selection event before the two incipient ecotypes had sufficiently diverged from one another.

Recombination may, in some cases, prevent a potentially speciation-quashing adaptive mutation from extinguishing another ecotype. If the adaptive mutation from one ecotype can recombine into another ecotype, the first ecotype may lose its advantage. In our "adapt globally, act locally" model of periodic selection,[30] the domain of competitive superiority of an adaptive mutant (i.e., the cell) is limited to its own ecotype, as I have described, but the adaptive mutation (i.e., the allele) can be recombined into other ecotypes. Upon transfer into another ecotype, an adaptive mutation precipitates a local periodic selection event within the recipient ecotype (Fig. 4). In this model, the chromosomal region near the adaptive mutation can be homogenized across different ecotypes, while the divergence elsewhere in the genome is unaffected.[30]

Note that a nascent ecotype is vulnerable to extinction by a parental ecotype only if its resource base is a subset of the parental ecotype's. When an ecotype utilizes at least one resource not used by the parental ecotype, it is then invulnerable to that ecotype's periodic selection events.[31] If a hypothesis by Lawrence[32,33] is correct, nascent ecotypes may readily acquire novel resources required to escape all periodic selection from their parental ecotype.

Lawrence[32,33] has argued that nearly all ecological divergence is precipitated by horizontal transfer, in which a recipient acquires a novel (heterologous) gene locus or operon from another species. By granting an entirely new metabolic function, heterologous gene transfer has the potential to endow a strain with a new resource base that is not shared with the parental ecotype. In this case, the horizontal transfer immediately places the nascent ecotype out of range of any periodic selection emanating from the parental ecotype.

In summary, ecological diversity in the bacterial world appears to be determined by three kinds of genetic changes (either mutations or recombination events) (Fig. 3). First, there are niche-invasion mutations (or recombination events), which allow the new genotype and its descendants to utilize a new set of resources and thereby escape periodic selection from the parental ecotype. Second, there are periodic selection mutations (or recombination events), which purge the diversity within a single ecotype; these tend to make ecotypes more distinct, since they purge the diversity within but not between ecotypes. Finally, there may be speciation-quashing mutations (or recombination events), whereby one ecotype can extinguish another. It will be interesting to quantify the rates at which these three kinds of mutations and recombination events occur, using microcosm evolution experiments that have been developed to investigate adaptive radiation in bacteria.[34-36]

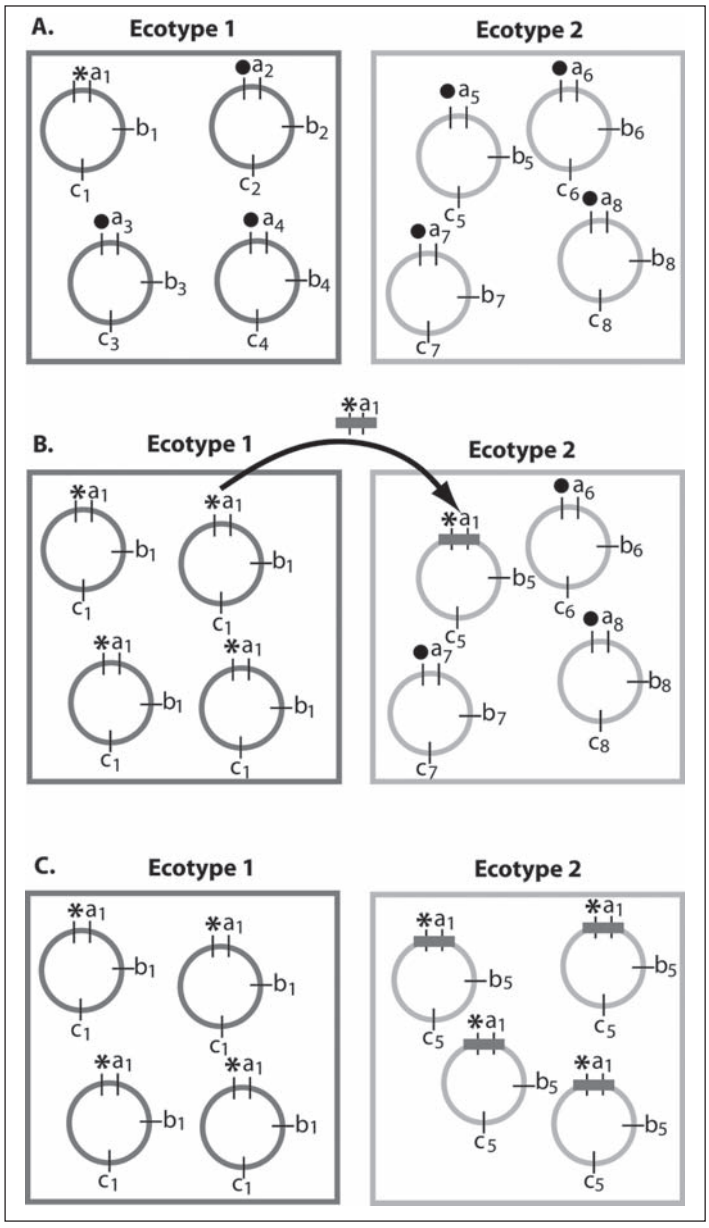

Figure 4. The "adapt globally, act locally" model of periodic selection. The domain of competitive superiority of an adaptive mutant (the cell) is the ecotype, as before, but the adaptive mutation (the allele) confers higher fitness to any individual in a variety of ecotypes. A) An adaptive mutation (represented by an asterisk) occurs in Ecotype 1. The disfavored, previously existing alleles are represented by black circles. B) The mutation sweeps the diversity within its own ecotype and then is transferred into Ecotype 2. C) The adaptive mutation now precipitates a periodic selection event within Ecotype 2. Each periodic selection event erases diversity genome-wide *within an ecotype*, but diversity between ecotypes is homogenized only in the chromosomal region closely linked to the adaptive mutation.

Periodic Selection and Discovery of Bacterial Ecotypes

Discovery of Ecotypes As Sequence Clusters

By purging diversity within but not between ecotypes, periodic selection provides a rationale for discovery of bacterial diversity. Given enough time, each bacterial ecotype is expected to be identifiable as a sequence cluster, distinct from all closely related ecotypes. In addition, each ecotype is expected to be identifiable as a monophyletic group in a phylogeny based on DNA sequence data.[26]

A phylogenetic perspective explains the predicted correspondence between ecotypes and DNA sequence clusters.[26] Suppose we begin with a single ecotype, and then one cell within the ecotype evolves new ecological properties and thereby founds a new ecotype. At this time the new ecotype appears in a phylogeny as if it is just one more lineage within the parental ecotype. However, the next adaptive mutant causing periodic selection within the parental ecotype will eliminate all other lineages within that ecotype, but will leave diversity within the nascent ecotype untouched. Likewise, periodic selection within the new ecotype will purge diversity within that ecotype but not within the parental ecotype. Recurrent selective sweeps within each lineage will result in long sequence distances on the phylogeny between each ecotype's contemporary diversity and the most recent ancestor shared by the two ecotypes (i.e., sequence distances will be much greater between than within ecotypes). Thus, each ecotype will eventually be discernible as a distinct sequence cluster and as a monophyletic group.

The Challenge of "Geotypes"

Care must be taken when using any sequence-based method to infer ecotypes. Geographically isolated populations that are members of the same ecotype could diverge into separate sequence clusters. In this case, an adaptive mutant from one geographic region is not given the chance to compete with populations from other regions, so sequence divergence between geographically isolated populations of the same ecotype could proceed indefinitely. Papke and Ward (personal communication) have argued that many bacterial taxa lack the means for worldwide travel, and so are expected to diverge into discrete clusters through geographic distance alone; they have termed the geographically based clusters within a single ecotype as "geotypes." As is the case for systematics of any organism, geography-associated sequence clusters of bacteria are difficult to interpret.

It is sometimes difficult to rule out the geotype hypothesis even when bacterial sequence clusters are sympatric. This may be the case when previously allopatric geotypes have only recently become sympatric, and have not yet had time for a periodic selection event to purge diversity throughout the ecotype. We will address this issue in the final section of the paper.

Discovery of Ecotypes As Star Clades

Another issue remains. A sequence-based phylogeny from almost any named bacterial species reveals a hierarchy of clusters, subclusters, and sub-subclusters. This raises the possibility that a typical named bacterial species may contain many cryptic and uncharacterized ecotypes, each corresponding to a small subcluster. The challenge is to determine which level of subcluster corresponds to ecotypes.

The peculiar dynamics of bacteria provide a method for identifying ecotypes based on sequence data.[22] Our "Star" approach assumes that the sequence diversity within an ecotype is constrained largely by periodic selection and much less by genetic drift, an assumption I will return to later.

Consider the consequences of periodic selection on the phylogeny of strains from the same ecotype. Nearly all stains randomly sampled from an ecotype should trace their ancestries

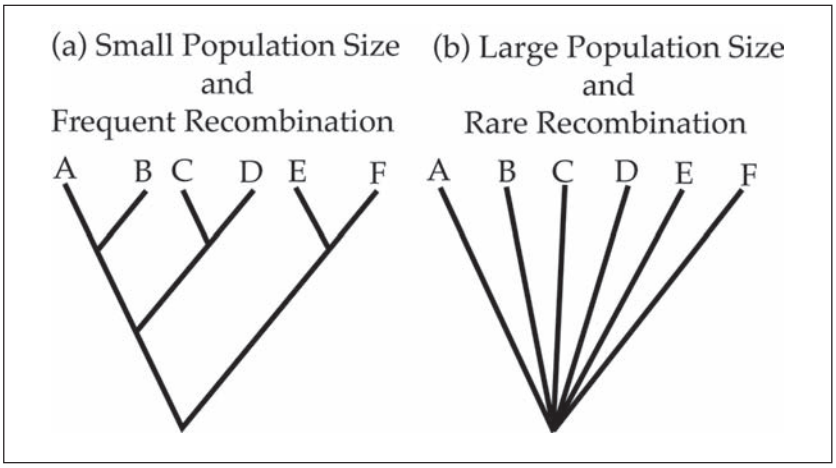

Figure 5. The phylogenetic signatures of ecotypes whose diversity is controlled by periodic selection versus genetic drift. a) In a population of small size, genetic drift causes coalescence of many pairs of lineages. Moreover, if recombination is frequent, there is no opportunity for genome-wide purging of diversity. Consequently, the phylogeny has many nodes. b) In a bacterial population, characterized by large population size and rare recombination, the population's phylogeny is expected to resemble a star. Following periodic selection, each strain traces its ancestry directly back to the adaptive mutant that precipitated the periodic selection event. In addition, population sizes are too large for genetic drift to create coalescences between pairs of strains with appreciable frequency.

directly back to the adaptive mutant that caused and survived the last selective sweep. Thus, the phylogeny of an ecotype should be consistent with a star clade, with only one ancestral node, such that all members of an ecotype are equally closely related to one another (Fig. 5B). In contrast, a population whose sequence diversity is limited by genetic drift will have a phylogeny with many nodes (Fig. 5A).

In an asexual ecotype, a sequence-based phylogeny would yield a perfect star clade, with only minor exceptions due to homoplasy. However, in a bacterial ecotype subject to modest rates of recombination, particularly with other ecotypes, the sequence-based phylogeny can deviate significantly from a perfect star. We have developed a computer simulation to determine how closely a sequence-based phylogeny of strains from the same ecotype should resemble a star clade. Taking into account the taxon's mutation and recombination parameters, the Star simulation determines the likelihood that the phylogeny of strains from a single ecotype would have only one significant node (i.e., a perfect star), versus two, three, four, or more significant nodes (significance determined by 95% bootstrap support).[22]

We found that for *S. aureus*, which is among the most clonal of bacteria, an ecotype's phylogeny should almost never have more than one node.[22] In the case of *N. meningitidis*, which is among the most frequently recombining bacteria, the phylogeny of an ecotype is expected to have one or two significant nodes, but almost never three or more. Accordingly, we may tentatively identify ecotypes of *N. meningitidis* as the largest clades that contain up to two significant nodes.

While Star produces a theory-based criterion for testing whether a set of strains belong to the same ecotype, this approach does not help us choose the groups of strains to be tested for membership within an ecotype. As I have previously shown,[22] the Multilocus Sequence Typing method (MLST) developed by B. Spratt and coworkers[37] produces reasonable hypotheses for demarcating strains of a named species into ecotypes.

Discovery of Ecotypes through Multilocus Sequence Typing

In MLST, partial sequences (450 bp) of seven gene loci that produce housekeeping proteins are surveyed. The evolutionary distance between strains is quantified in MLST as the number of loci that are different, whether by substitution of a single nucleotide or a swath of nucleotides (possibly due to a recombination event). Strains are then classified into "clonal complexes": all strains that are identical with a particular strain at five or more loci (in some cases, six or more loci) are deemed members of a clonal complex. E. Feil has developed the "Burst" computer algorithm for assigning strains into clonal complexes according to criteria set by the user (web site: www.mlst.net).

The clonal complexes defined by MLST correspond remarkably well to ecologically distinct groups, even in taxa such as *N. meningitidis* where recombination is unusually high. I have hypothesized that the clonal complexes identified by MLST are ecotypes.[5,22] Because periodic selection is recurrently purging the diversity within an ecotype, ecotypes are expected to accumulate only a limited level of sequence diversity between periodic selection events, depending on the rates of mutation and recombination (which generate variation) and the time between periodic selection events. We may speculate that ecotypes typically have only enough time between selective sweeps for a given strain to accumulate divergence at one or two loci out of seven, on average, whether by mutation or recombination. This yields the 5/7 and 6/7 criteria used in MLST. In general, one would expect that frequently recombining bacteria (in which a locus is ten times more likely to be struck by a recombination than a mutation event) would diverge at more loci between periodic selection events, compared to rarely recombining bacteria, where nearly all divergence accumulates simply through mutation.

Because MLST's 5/7 and 6/7 criteria are intuitively based, we should test whether MLST's clonal complexes do indeed correspond to ecotypes. The Star algorithm can test whether the clonal complexes identified with MLST have phylogenies consistent with ecotypes, taking into account the recombination and mutation parameters estimated for the particular taxon.

I have tested whether the strains of each of the ten clonal complexes found within *N. meningitidis* are consistent with the Star simulation's expectations for a single ecotype.[22] It turns out that the phylogenies of all but one of the ten MLST clonal complexes within *N. meningitidis* contain one or two nodes, as expected given *N. meningitidis*'s recombination parameters. Similarly, all but three of the 26 clonal complexes within *S. aureus* are consistent with the expectation for a single ecotype (i.e., containing no more than one significant node). The three exceptional clonal complexes, when pooled together, contain only one significant node among them, suggesting that they are members of the same ecotype. Taking into account the rarity of recombination in *S. aureus*, perhaps the criterion for inclusion within a clonal complex in this taxon should be 6/7 instead of 5/7 identical loci. It will be interesting to use the Star approach to calibrate the Burst criterion.

In summary, Star demonstrates that the clonal complexes yielded by MLST have phylogenies consistent with ecotypes, at least within *S. aureus* and *N. meningitidis*. The clonal complexes produced by MLST do indeed yield reliable hypotheses about the membership of ecotypes, and these hypotheses can be tested using Star.

It is striking that each named species studied by MLST has so many clonal complexes.[11,12] If each of these clonal complexes can be shown definitively to be a separate ecotype, each with the universal properties of species, a named bacterial species may actually be more like a genus than a species.[22]

We should regard the ecotypes identified by sequence-based approaches as only putative until each ecotype can be shown to be ecologically distinct. Ideally, we should also demonstrate that each group has undergone its own private periodic selection events. This is because two putative ecotypes that are only slightly different ecologically may be subject to extinction by one another's periodic selection events. To show that each putative ecotype has already undergone

one or more separate periodic selection events would bolster our claim that the clusters and clonal complexes we identify are actually distinct ecotypes. In the next section, I outline a sequence-based method for demonstrating that putative ecotypes have undergone their own, private periodic selection events.

Has Periodic Selection Occurred in Nature?

Periodic Selection Is Inevitable at Bacterial Recombination Rates

We have seen that the rare recombination typical of bacteria is not sufficient to preserve a bacterial population's genetic diversity (Fig. 2). Each adaptive mutation that moves to fixation will eliminate nearly all of the sequence diversity, depending on the recombination rate and the intensity of selection (and to a lesser extent, the population size). Given the inevitability of mutational improvements over time, should we not expect to see recurrent purges of diversity?

Notley-McRobb and Ferenci[2] have argued that recurrent adaptive mutations will not necessarily purge diversity within a rarely sexual population of bacteria. This may be the case if, in each bout of selection, there are many independently derived adaptive mutations, each stemming from a different genetic background.[38] For example, Notley-McRobb and Ferenci[2] found that as many as 13 adaptive mutant alleles at the *mlc* locus swept simultaneously through an experimental population of *E. coli*. Thus, instead of one genetic background sweeping through the population, there would be 13, diminishing the diversity-purging effect of periodic selection.

Nevertheless, I believe we should not conclude that periodic selection is generally ineffective at purging diversity. First, even if the periodic selection events in Notley-McRobb and Ferenci's[2] study are typical of nature, periodic selection should still purge population diversity, albeit at a slower rate. Note that strong selection favoring a single adaptive mutation sweeps all but 0.2% of the diversity from a population, even under high recombination rates. If there were instead ten adaptive mutations sweeping the population simultaneously, the fraction of diversity saved would be increased by only a factor of ten, in this case to 2% of the population's original diversity. So, whether one or several adaptive mutations sweep through each periodic selection event, recurrent sweeps will indeed limit population diversity to very low levels.

Moreover, we cannot expect every selective sweep to be driven by an ensemble of equally fit adaptive mutations. This will be the case only when a population is placed in a new environment and many equally good mutations can accommodate the environmental change. For example, we know that the glucose-limiting environment of Notley-McRobb and Ferenci's experiment favors changes in the regulatory loci *mlc*, *mglD*, and *malT*, and many mutations at these loci appear equally good at adjusting the cells to this environment.

On a global or even regional scale, environmental change is not likely to cause selective sweeps across the entire geographical range of an ecotype, unless the environmental change is global. Because different local populations are likely to see environmental changes in different directions at a given moment, the adaptive mutations that accommodate environmental change are not the ones that would sweep an entire ecotype in all its localities.

What are the adaptive mutations most likely to sweep an ecotype throughout all its diverse habitats? These are mutations (or recombination events) that bring about a novel and generally useful adaptation, which improves the fitness of the ecotype throughout its range of living situations. For example, these might be overall improvements in efficiency, which are adaptive in any circumstance. All the easily accessible mutations of this nature have already been obtained (e.g., all single-nucleotide substitutions are immediately accessible owing to the large population sizes of bacteria); those that remain are much less frequent changes that involve two or more simultaneous mutations (e.g., where none is adaptive without the others), or perhaps recombination with other species. In contrast to the case for a change in

environment, where many mutations can independently and simultaneously rise to the occasion, the wait for an innovation is rewarded by only a single rare event. And when this event occurs in a rarely recombining ecotype of bacteria, it will purge the diversity, ecotype-wide, as predicted in Figure 2.

Periodic Selection Explains the Small Effective Population Sizes of Bacteria

A named species of bacteria typically has only a modest level of DNA sequence variation in protein-coding genes, ranging from less than 0.01 to 0.05 (average pair wise sequence divergence). If sequence diversity were limited by genetic drift, a typical species-wide diversity level (e.g., 0.02 for *N. meningitidis*) would correspond to an effective population size (N_e) of 2.2×10^7 (assuming the whole named species to be a single ecotype).[21] Given the enormous census sizes of bacteria in nature, this estimate of N_e appears absurdly low. However, periodic selection can reasonably explain the low levels of sequence diversity typical for bacteria. Periodic selection occurring once in N_e generations will yield the same amount of diversity as pure drift would in a population of size N_e. Thus, periodic selection occurring once in 2.2×10^7 generations would yield the diversity levels typical of named species (e.g., *N. meningitidis*). Periodic selection need not occur often to constrain ecotype-wide diversity to modest levels.

As I have argued earlier, the range of periodic selection events is more likely to correspond to MLST's clonal complexes than to named species. In this case, a much lower frequency of ecotype-wide periodic selection would be required to explain observed diversity levels. For example, the average sequence diversity within *N. meningitidis*'s clonal complexes (0.005) would require ecotype-wide periodic selection occurring once in 6×10^6 generations.

Identification of Periodic Selection Events

A survey of sequence diversity in *E. coli*[39] has provided the clearest evidence for a periodic selection event in nature. Most genes in the survey corroborated the population structure already known from allozyme data: strains fell into four major sequence clusters. However, within one gene region, near *gapA*, all strains were anomalously homogeneous in sequence. These results were interpreted as evidence for a selective sweep throughout *E. coli*, driven by an adaptive mutation in the *gapA* region.

This interpretation would be entirely appropriate if *E. coli* were a highly sexual species. In the case of animals and plants, the diversity-purging effect of natural selection is limited to the chromosomal region near the adaptive mutation, where recombination with the adaptive mutation is infrequent. In bacteria, however, recombination between any two genes, regardless of their distance on the chromosome, is extremely rare. Therefore, if all of *E. coli* were a single ecotype, and the sequence homogeneity around *gapA* were caused by an ecotype-wide purging of diversity, we would expect all of *E. coli* to be purged of diversity over the entire chromosome.

Jacek Majewski and I[30] previously proposed the adapt globally, act locally model to explain anomalous homogeneity around a small chromosomal region, as seen for *gapA* in *E. coli*. Because *E. coli* forms four major sequence clusters, as well as many smaller subclusters, we may tentatively conclude that *E. coli* contains several ecotypes. We proposed that the adaptive mutation around *gapA* was generally useful for all of the ecotypes of *E. coli*, and that the allele was passed between ecotypes, precipitating a periodic selection event within each (Fig. 4). Thus, for genes closely linked to the adaptive mutation, there would be nearly total purging of diversity both within and between ecotypes, but for genes not linked to the adaptive mutation, selection would purge only the diversity within ecotypes. Whenever a small chromosomal segment is homogenized across strains that otherwise form distinct clusters, a generally useful adaptive mutation is likely to have passed from ecotype to ecotype, causing local periodic selection in each.

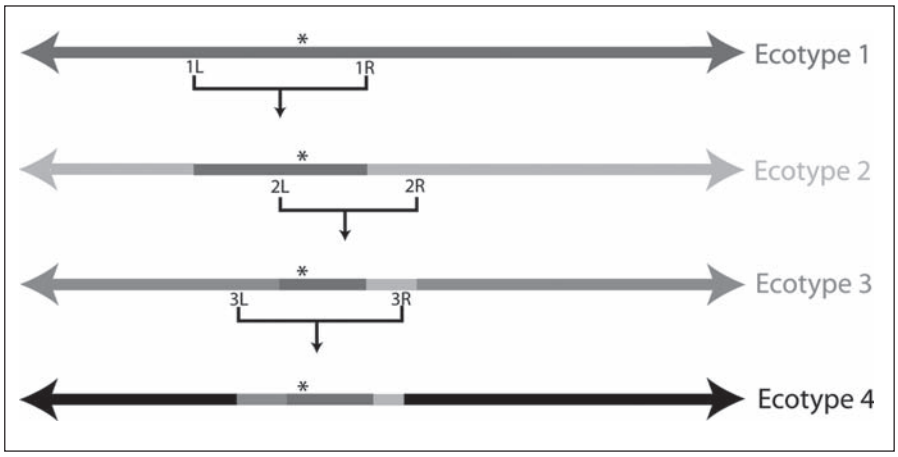

Figure 6. In the adapt globally, act locally model, the region that is homogenized is expected to differ between each pair of ecotypes. The adaptive mutation (indicated by an asterisk) originally occurs in Ecotype 1. After the selective sweep, a small region of chromosome around the adaptive mutation (between 1L and 1R) enters Ecotype 2, causing a selective sweep in that ecotype. Then, a segment around the adaptive mutation (between 2L and 2R) enters Ecotype 3, and causes a selective sweep there, and so on. The source of DNA along the chromosome is indicated by shade. The entire ensemble of ecotypes becomes homogenized for the sequence near the adaptive mutation, but the boundaries of the homogeneous region differ for each pair of ecotypes. For example, Ecotypes 1 and 2 are identical between 1L and 1R, while Ecotypes 2 and 4 are identical between 2L and 3R.

Beyond providing evidence for periodic selection, the sequence pattern found by Guttman and Dykhuizen[39] can provide additional evidence that sequence clusters correspond to ecotypes. Recall that the clusters we discover may correspond either to ecotypes or to geotypes, which are populations of the same ecotype with a history of geographic isolation. This issue can be resolved when we find evidence of the Guttman-Dykhuizen pattern.

Different sequence clusters can correspond to geotypes within the same ecotype only if there has not been an opportunity for periodic selection to sweep through all of the clusters. If we can show that these clusters have survived as distinct groups through a periodic selection event, we can rule out the geotype hypothesis. This is indeed the case for the various clusters within *E. coli*. The homogenization of the *gapA* region across the various major clusters in *E. coli* shows that these clusters have maintained their distinctness, even as an adaptive mutation has caused periodic selection events within each cluster.

The Guttman-Dykhuizen pattern can provide further evidence of multiple ecotypes. We should expect that the adaptive allele driving the periodic selections in all of the ecotypes is passed to each ecotype in a separate recombination event. Therefore, the region that is homogenized should be somewhat different for each pair of ecotypes, reflecting the junctions of the recombination events that transferred the adaptive mutation across ecotypes (Fig. 6). We may thus predict that if the sequence clusters correspond to ecotypes, the junctions of the homogeneous region will be unique for each pair of sequence clusters.

Guttman and Dykhuizen's[39] discovery of a periodic selection event was based on a serendipitous choice of loci to survey, but today comparisons of whole genomes should provide ample opportunities for genome-wide screening of periodic selection events driven by "adapt globally, act locally" mutations. Discovery of these periodic selection events would allow us to confirm that the many sequence clusters found within named species are distinct ecotypes.

Acknowledgements

I am grateful to Michael Dehn for help in clarifying the manuscript. This work was supported by National Science Foundation grants DEB-9815576 and EF-0328698 and research funds from Wesleyan University.

References

1. Atwood KC, Schneider LK, Ryan FJ. Periodic selection in Escherichia coli. Proc Natl Acad Sci USA 1951; 37:146-155.
2. Notley-McRobb L, Ferenci T. Experimental analysis of molecular events during mutational periodic selections in bacterial evolution. Genetics 2000; 156(4):1493-1501.
3. Maynard Smith J, Smith N, O'Rourke M et al. How clonal are bacteria? Proc Natl Acad Sci USA 1993.
4. Cohan FM. Clonal structure: An overview. In: Pagel M, ed. Encyclopedia of Evolution. New York: Oxford University Press, 2002a:158-161.
5. Cohan FM. Population structure and clonality of bacteria. In: Pagel M, ed. Encyclopedia of Evolution. New York: Oxford University Press, 2002b:161-163.
6. Ochman H, Lawrence JG, Groisman EA. Lateral gene transfer and the nature of bacterial innovation. Nature 2000; 405(6784):299-304.
7. Posada D. Evaluation of methods for detecting recombination from DNA sequences: Empirical data. Mol Biol Evol 2002; 19(5):708-717.
8. Maynard Smith J, Smith NH. Detecting recombination from gene trees. Mol Biol Evol 1998; 15(5):590-599.
9. Hey J, Wakeley J. A coalescent estimator of the population recombination rate. Genetics 1997; 145(3):833-846.
10. Hudson RR. Estimating the recombination parameter of a finite population model without selection. Genet Res 1987; 50(3):245-250.
11. Feil EJ, Maiden MC, Achtman M et al. The relative contributions of recombination and mutation to the divergence of clones of Neisseria meningitidis. Mol Biol Evol 1999; 16(11):1496-1502.
12. Feil EJ, Smith JM, Enright MC et al. Estimating recombinational parameters in Streptococcus pneumoniae from multilocus sequence typing data. Genetics 2000; 154(4):1439-1450.
13. Guttman DS, Dykhuizen DE. Clonal divergence in Escherichia coli as a result of recombination, not mutation. Science 1994a; 266(5189):1380-1383.
14. Feil EJ, Spratt BG. Recombination and the population structures of bacterial pathogens. Annu Rev Microbiol 2001; 55:561-590.
15. Suerbaum S, Maynard Smith J, Bapumia K et al. Free recombination within Helicobacter pylori. Proc Natl Acad Sci USA 1998; 95:12619-12624.
16. Drake JW. A constant rate of spontaneous mutation in DNA-based microbes. Proc Natl Acad Sci USA 1991; 88(16):7160-7164.
17. Duncan KE, Istock CA, Graham JB et al. Genetic exchange between Bacillus subtilis and Bacillus licheniformis: Variable hybrid stability and the nature of species. Evolution 1989; 43:1585-1609.
18. Cohan FM. Bacterial species and speciation. Syst Biol 2001; 50:513-524.
19. Lawrence JG, Ochman H. Molecular archaeology of the Escherichia coli genome. Proc Natl Acad Sci USA 1998; 95(16):9413-9417.
20. Braverman JM, Hudson RR, Kaplan NL et al. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics 1995; 140(2):783-796.
21. Cohan FM. The effects of rare but promiscuous genetic exchange on evolutionary divergence in prokaryotes. Am Naturalist 1994; 143:965-986.
22. Cohan FM. What are bacterial species? Annual Review of Microbiology 2002c; 56:457-487.
23. Templeton A. The meaning of species and speciation: A genetic perspective. In: Otte D, Endler J, eds. Speciation and its Consequences. Sunderland MA: Sinauer Assoc, 1989.
24. Simpson G. Principles of Animal Taxonomy. New York: Columbia Univ Press, 1961.
25. Wiley E. The evolutionary species concept reconsidered. Syst Zool 1978; 27:17-26.
26. Palys T, Nakamura LK, Cohan FM. Discovery and classification of ecological diversity in the bacterial world: The role of DNA sequence data. Int J Syst Bacteriol 1997; 47(4):1145-1156.

27. Sokal R, Crovello T. The biological species concept: A critical evaluation. Am Nat 1970; 104:127-153.
28. Mallet J. A species definition for the modern synthesis. Trends in Ecology & Evolution 1995; 10:294-299.
29. VanValen L. Ecological species, multispecies, and oaks. Taxon 1976; 25:233-239.
30. Majewski J, Cohan FM. Adapt globally, act locally: The effect of selective sweeps on bacterial sequence diversity. Genetics 1999; 152(4):1459-1474.
31. Holt RD. On the relationship between niche overlap and competition: The effect of incommensurable niche dimensions. Oikos 1987; 48:110-114.
32. Lawrence J. Gene Transfer in Bacteria: Speciation without Species? Theor Popul Biol 2002; 61(4):449.
33. Lawrence J. Catalyzing Bacterial Speciation: Correlating Lateral Transfer with Genetic Headroom. Syst Biol 2001; 50(4):479-496.
34. Rainey PB, Travisano M. Adaptive radiation in a heterogeneous environment. Nature 1998; 394(6688):69-72.
35. Treves DS, Manning S, Adams J. Repeated evolution of an acetate-crossfeeding polymorphism in long-term populations of Escherichia coli. Mol Biol Evol 1998; 15(7):789-797.
36. Imhof M, Schlotterer C. Fitness effects of advantageous mutations in evolving Escherichia coli populations. Proc Natl Acad Sci USA 2001; 98(3):1113-1117.
37. Maiden MC, Bygraves JA, Feil E et al. Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci USA 1998; 95(6):3140-3145.
38. Korona R. Genetic divergence and fitness convergence under uniform selection in experimental populations of bacteria. Genetics 1996; 143(2):637-644.
39. Guttman DS, Dykhuizen DE. Detecting selective sweeps in naturally occurring Escherichia coli. Genetics 1994b; 138(4):993-1003.

Chapter 8

Distribution and Abundance of Polymorphism in the Malaria Genome

Stephen M. Rich

Plasmodium falciparum is the most deadly of the four human malaria parasites, causing as many as 500 million malaria cases per year and more than 2 million deaths.[1] Despite more than a century of biomedical research and unprecedented (indeed, unsurpassed) measures of international collaboration to eradicate the disease, the situation only seems to be worsening as drug-resistant parasites come to dominate the landscape. Indeed, *P. falciparum* has demonstrated remarkable adaptive potential in overcoming every effort to thwart its transmission. Novel strategies are currently in development. These include the innovation of new therapeutic modalities,[2,3] development of protective vaccines,[4-6] and efforts to develop refractory mosquito vectors.[7,8] Choosing the most effective means of reducing malaria transmission will require careful consideration of the parasite's ability to circumvent targeted interventions of its transmission cycle. For example, objective criteria should be established for prioritizing among the 40+ vaccines currently in development and for assessing the sustainability of their protection. Accordingly, it is crucial that we discern the evolutionary processes that have facilitated the persistent association of the parasite and its human host. In short, it is imperative to determine how genetic variation within and among extant *P. falciparum* population actuates to become the parasites' adaptive response to vaccine and drug pressures.

Malariologists have long recognized the importance of genetic variation and their attempts to quantify it predate the widespread availability of nucleotide-based assays of genetic diversity. Comparative serological studies and other protein-based methodologies demonstrated that *P. falciparum* populations are comprised of antigenically diverse sets of strains,[9-11] to the extent that even parasite isolates taken from individual patients may harbor multiple antigen types. The advent of PCR technologies in the late 1980's made it possible to expand upon earlier studies by quantifying genetic variation directly from nucleotide sequences. These molecular studies focused initially on the genes that encoded antigenic determinants and ribosomal subunits.

With the availability of nucleotide sequences of *P. falciparum* and other malaria parasites, it became possible to estimate phylogenetic relatedness of the members of the genus *Plasmodium*. Hence it was determined that, in stark contrast to the situation in the other human malaria parasites, *P. falciparum* has shared a parallel evolutionary trajectory with its chimpanzee counterpart, *P. reichenowi*.[12,13] The time of divergence between these two *Plasmodium* species was estimated at 5-7 million years (My) ago, which is roughly consistent with the time of divergence between the two host species, human and chimpanzee.[14] The parsimonious interpretation is that *P. falciparum* is an ancient human parasite associated with our ancestors at least

Selective Sweep, edited by Dmitry Nurminsky. ©2005 Eurekah.com and Kluwer Academic/Plenum Publishers.

since the divergence of the hominids from the great apes, and that the divergence of *P. falciparum* and *P. reichenowi* was concurrent with the divergence of their host species, humans and chimps.

Since *P. falciparum* shares a long evolutionary history with its human host, some investigators began to hypothesize that allelic forms of *P. falciparum* antigenic determinants may also be quite old. Based on the observed diversity of genes encoding surface protein molecules, Hughes surmised that some antigenic variability may be as old as 35 million years, or half as old as the *Plasmodium* genus.[15,16] This situation would be analogous to that found in human *Hla* loci, where extant alleles are extremely divergent from one another due to the extreme age of the variants which predates the split of humans from other hominoid primates.[17] However, replacement polymorphisms in antigenic genes such as those in Hughes' study, are usually under strong diversifying selection imposed by the host's immune system.[18] Therefore their rates of evolution are likely to be quite erratic and may not yield accurate estimates of the age of species.[19] More accurate measures of allelic age would be afforded by analysis of synonymous polymorphisms (SNP's), which do not alter the encoded amino acid and so are thought to be evolving by largely neutral mechanisms. Nucleotide substitutions at these sites occur at a steady rate through geological time periods, as a function of the mutation rate and elapsed time. Mutation rates can be obtained empirically by counting differences among gene sequences from species for which the divergence time is known, such as the case for *P. falciparum* and *P. reichenowi*.

Determining the age of allelic variation is important for identifying how organisms utilize genetic resources to cope with the environment. A practical goal would be to limit genetic variation among parasite populations and hence reduce their adaptive potential, thus it is necessary to identify the means by which this variation is generated and maintained. For example, in the case described above for *Hla* alleles, genetic variation is extremely old and humans have maintained the adaptive potential of these variants by maintaining a balanced distribution of alleles among individuals and populations. Alternatively, the variable surface glycoprotein (*Vsg*) genes of *Trypanosoma cruzi* parasites have allelic variants that are completely ephemeral and are being constantly regenerated by duplicative transposition of component sequences. In the case of the *Hla*, alleles are not easily lost since they are distributed among many individuals' genomes, however, loss of an allele is extremely detrimental since replacements arise slowly. Alternatively, *Vsg* alleles disappear frequently but are quickly replaced by equivalents, hence adaptive potential is more closely linked to diversity generating mechanisms rather than maintenance of individual variants. Discriminating between these two alternative scenarios requires careful consideration of the genomic context in which the genes of interest are found. For example, in order to determine whether *P. falciparum* antigenic alleles are ancient, it is necessary to examine their polymorphisms relative to those loci representative of the balance of the genome and particularly among genetic loci not under immune selection.

With this in mind, we sought to determine the age of extant distributions of *P. falciparum* by quantifying SNP's among isolates from global locations.[20] When we excluded those loci known to be under strong selection, i.e., antigenic determinants, we found a marked paucity of SNP's. Indeed, among the > 30,000 synonymous sites distributed among 10 genetic loci examined from dozens of parasite isolates collected on 4 different continents, we did not find a single SNP. Based on these observations we estimated that the current distribution of *P. falciparum* throughout the world's tropical regions is derived from a small ancestral population in the very recent past. We referred to this conclusion as the Malaria's Eve hypothesis, and we estimated the upper confidence interval of the age of this recent ancestry at 8,000-60,000 years based on different estimates of the SNP mutation rate.[19,20]

In the few short years following our first report of Malaria's Eve, the issue has created a contentious debate. Our initial conclusion was based on nucleotide sequences that were then available from GenBank, and the only criteria for inclusion of genes in our dataset was that

they had to be void of repetitive DNA sequences and show no evidence of being under positive selection. At that time (1998), the amount of sequence data available for the species was rather limited, but since then this dataset has grown exponentially, culminating in the complete genome sequence of *P. falciparum* published in 2002.[21] This spurred other investigators to carefully scrutinize the Malaria's Eve hypothesis.

One of these studies entailed a large scale sequencing survey of 25 introns located on the second chromosome, from eight *P. falciparum* isolates collected among global sites.[22] The findings of this study confirmed our previous result: there is an extreme scarcity of silent site polymorphism among extant distributions of *P. falciparum*. Among some 32,000 nucleotide sites examined, Volkman et al found only 3 silent single nucleotide polymorphisms (SNP's) and concluded that the age of Malaria's Eve was somewhere between 3,200 and 7,700 years.[22,23]

Conway et al[24] have presented further evidence in support of Malaria's Eve based on analysis of the *P. falciparum* mitochondrial genome. They examined the entire mitochondrial DNA (mtDNA) sequence of *P. falciparum* isolates originating from Africa (NF54), Brazil (7G8), and Thailand (K1 and T9/96), as well as the chimpanzee parasite, *P. reichenowi*. Alignment of the four complete mtDNA sequences (5,965 bp) showed that 139 sites contain fixed differences between *falciparum* and *reichenowi*, whereas only 4 sites were polymorphic within *falciparum*. The corresponding estimates of divergence (K, between *P. reichenowi* and *P. falciparum*) and diversity (π, within *P. falciparum* strains), are 0.1201 and 0.0004, respectively. In short, divergence in *mtDNA* sequence between the two species is 300-fold greater than the diversity within the global *P. falciparum* population. If we use the *rDNA*-derived estimate of 8 million years as divergence time between *P. falciparum* and *P. reichenowi*, then the estimated origin of the *P. falciparum* mtDNA lineages is 26,667 years (i.e., 8 million/ 300), which corresponds quite well with our estimate based on 10 nuclear genes.[20] In a subsequent survey of a total of 104 isolates from Africa (n=73), Southeast Asia (n=11), and South America (n=20); Conway et al[24] determined that the extant global population of *P. falciparum* is derived from three mitochondrial lineages that started in Africa, and migrated subsequently (and independently) to South America and Southeast Asia. Each mitochondrial lineage is identified by a unique arrangement of the 4 polymorphic *mtDNA* nucleotide sites.

Arguments against the Malaria's Eve hypothesis come in two forms. The first argument is that the loci chosen in the studies described above are a biased sample and do not reflect the levels of polymorphism in the genome as a whole. The second counterargument concedes that nucleotide polymorphisms are scarce, however contends that this paucity is not attributable to recent origin, but rather reflects strong selection pressure against the occurrence of synonymous SNP's.

One such study reports an "ancient" origin of *P. falciparum* based on a survey of sequences available from the GenBank database.[25] It should be pointed out that some of these GenBank sequences are compiled from a variety of sources and many of the entries may contain sequencing errors associated with misincorporation of nucleotides by Taq polymerase during the PCR amplification of alleles. Moreover, some of the sequences included in the paper were not carefully examined, and the comparisons include multiple nucleotide sequences from a single clone derived in different laboratories. For example, GenBank entries AF239801 and AF282975 are both *falcipain-2* sequences from *P. falciparum* clone W2. Regardless of possible errors, the overwhelming message from their compiled data is that there is indeed a dearth of polymorphism. In fact, among the 23 loci examined, which comprised over 10,000 codons, only six contained synonymous SNP's in 4-fold degenerate codons. Nonetheless, Hughes and Verra conclude that time to most recent common ancestry of *P. falciparum* may be 300,000-400,000 years.

A more ambitious effort to quantify polymorphism in *P. falciparum* involved a survey of > 200 kb from the completed chromosome 3.[26] The authors reported 31 and 62 polymorphisms

among 80,415 noncoding and 192,400 synonymous nucleotide sites, respectively. Using the equation and mutation rates from our paper,[20] Mu et al estimated that the common ancestor to be between 102,000 and 177,000 years old. At this level of polymorphism, i.e., 62 of 192,400 (or 0.03%), the error rate in PCR and sequencing becomes relevant and bears great impact on estimates of recent ancestry. Mu et al[26] reamplified and resequenced 56 of the SNP containing regions and in this second pass found that 2 of the SNP's were in error (an error rate of ~4.0%). Because of this, the previously described paper by Volkman et al[22] incorporated highly redundant approach to assure integrity of the data. Volkman et al's methods involved meticulous bi-directional sequencing of 3 clones from each of 3 independent PCR amplifications, or an 18-fold redundancy.[23]

Another concern about calculation of the age of Malaria's Eve pertains to the estimation of mutation rates. The estimates used by Mu et al[26] are from a very small number of nucleotides (708 bp) compared between the rhoptry-associated protein gene of *P. falciparum* and *P. reichenowi*.[20] The neutral mutation rate may vary among chromosomal regions and its estimation is subject to sampling error. Even slight perturbations in its calculation will have exponential effects on estimation of age of the common ancestor. Reliable estimates of the mean age of the recent common ancestor are in the range of 4,000 to 180,000 years. While at first glance these differences of nearly two-orders of magnitude appear unsatisfactory, the differences are in fact quite small in light of the 5 MY age of the species dating back to its split with the chimpanzee parasite. This means that the global, extant distribution of *P. falciparum*, with its abundant diversity of antigens and drug resistance factors, originated in only a small fraction (at most ~3%) of the time since the origin of the species. This finding contrasts greatly with the previous estimates of some antigenic variation to the on order of 35 MY old.[16,19]

Despite discrepancies in the estimation of age of the Malaria's Eve common ancestry, it is clear that nucleotide polymorphisms are scarce in many portions of the *P. falciparum* genome.[23,27] A second criticism of the recent origins hypothesis concedes the paucity of synonymous site polymorphism but attributes this is due to constraints on the genome itself. One proposition is that the extreme AT content of the *P. falciparum* genome may suggest that some constraint is acting upon mutations that lead to unfavorable codon sequences.[28-30] As we have argued elsewhere, this does not seem to be the case, since in spite of AT content as high as 84% in third positions, there appears to be an equal proportion of A and T nucleotides in third positions of four-fold degenerate codons.[31,32] Moreover, the fact that synonymous substitutions are in evidence in the divergence between *P. falciparum* and *P. reichenowi* (which has a similarly extreme AT content), indicates that mutations can and do occur.[32]

Hartl et al[23] have pointed out that genomic constraints seem unlikely given the variability of microsatellite markers among introns, intergenic regions and in some cases, coding sequences.[22,33-35] Nonetheless, Forsdyke[36] has argued that the extreme conditions of the *P. falciparum* genome present a situation where selection for genomic composition exceeds the selection on the proteins encoded by these genes. The argument is leveled not so much against the Malaria's Eve hypothesis in particular, but rather the author attempts to refute the notion that neutral evolution is even possible. This warrants further discussion.

In an attempt to assign adaptive significance to the occurrence of a simple-repetitive sequence element (the Epstein-Barr Nuclear Antigen-1, *EBNA-1*) in the genome of the Epstein-Barr virus (EBV), Forsdyke[36] argues that the selective pressure for particular genomic content and/or arrangement supercedes the selection acting on encoded proteins (phenotype). The *EBNA-1* can be removed from the genome without any loss of function in the virus. Because EBV, like most viruses, tends to lose extraneous genetic elements nonessential to its survival, Forsdyke[36] maintains that the *EBNA-1* must have a function other than that typically assigned to genes, i.e., to encode message. To establish this fact, he has developed several descriptive parameters that are based on the nucleotide composition and secondary-folding

potential of nucleotide sequences. These parameters are termed as a potential "pressures" acting on the genome to maintain a particular configuration and/or composition. Forsdyke[36] tested whether the region in question has extraordinary values for the pressure parameters, and found that in the EBNA-1 region, there was an excessive skew in purine content (A and G) which would limit the potential for folding of the molecule and hence reduce recombination. The potential benefit of this situation is not explained and its biological relevance remains unclear.

The analysis of the EBV provided the analytical bases of Forsdyke's claim that *P. falciparum* is under pressure for reduced nucleotide polymorphism. He chose to examine the individual sequence content of two *P. falciparum* surface antigens, *Csp* and the merozoite surface protein-2 (*Msp-2*). As with the EBNA-1, he found that there was a high bias toward purines (primarily A in this case) and a strong potential for secondary folding within the repetitive regions of both *Msp-2* and *Csp*. The only conclusion drawn from this was that the high folding potential might enhance recombination in the repeat regions of both genes. The model is neither predictive nor explanatory, and even offers very little in the way of descriptive value. If it were demonstrated that these extraordinary regions had significantly less (or greater) numbers of SNP's, and that pressure parameters were predictive of this polymorphism, the author's claim may bear some relevancy. However, neither of these claims can be made particularly because the author chose to examine two of the most highly polymorphic loci known in *P. falciparum*. What is clear is that silent site polymorphisms are in evidence in nonfalciparum malaria species, and that synonymous substitutions have occurred in the divergence of *P. falciparum* and *P. reichenowi*. On this basis, we maintain that while substitutions may be constrained due to nucleotide composition and/or codon usage bias, these constraints do not explain the paucity of *P. falciparum* synonymous site variation. Therefore, the Malaria's Eve hypothesis remains the most likely explanation for this state of affairs.

In addition to the analyses of genetic polymorphisms data, there is independent information in support of the Malaria's Eve hypothesis. Sherman[37] notes the late introduction and low incidence of *falciparum* malaria in the Mediterranean region. Hippocrates (460-370 B.C.) describes quartan and tertian fevers, but there is no mention of severe malignant tertian fevers, which suggests that *P. falciparum* infections did not yet occur in classical Greece, as recently as 2,400 years ago. Interestingly, Tishkoff et al[38] traced the origin of malaria-resistant *G-6pd* genotypes in humans to the spread of agricultural societies some 5000 years ago. The recent origin of this mutation in humans suggests a similarly recent association with widespread exposure to the malaria parasite.

How can we account for a recent demographic sweep of *P. falciparum* across the globe, given its long-term association with the hominid lineage? One likely hypothesis is that human parasitism by *P. falciparum* has long been highly restricted geographically, and has dispersed throughout the Old World continents only within the last several thousand years, perhaps within the last 10,000 years, after the Neolithic revolution.[39,40] Three possible scenarios may explain this historically recent dispersion: (1) changes in human societies, (2) genetic changes in the host-parasite-vector association that have altered their compatibility, and (3) climatic changes that entailed demographic changes (migration, density, etc.) in the human host, the mosquito vectors, and/or the parasite.

The current, globally widespread distribution of *P. falciparum* from a limited original concentration probably in tropical Africa, may be attributed to changes in human living patterns – particularly the development of agricultural societies and urban centers that increased human population density.[37,40-45] Genetic changes that have increased the affinity within the parasite-vector-host system are also a possible explanation for a recent expansion, not mutually exclusive with the previous one. Coluzzi[40,41] has cogently argued that the worldwide distribution of *P. falciparum* is recent and has come about, in part, as a consequence of a recent

dramatic rise in vectorial capacity due to repeated speciation events in Africa of the most anthropophilic members of the species complexes of the *Anopheles gambiae* and *A. funestus* mosquito vectors. Biological processes implied by this account may have been associated with, and even dependent on the onset of agricultural societies in Africa (scenario 1) and climatic changes (scenario 3), specifically gradual increase in ambient temperatures after the Würm glaciation, so that about 6,000 years ago climatic conditions in the Mediterranean region and the Middle East made the spread of *P. falciparum* and its vectors beyond tropical Africa possible.[40,41,44,45] The three scenarios are likely interrelated. Once demographic and climatic conditions became suitable for propagation of *P. falciparum*, natural selection would have facilitated evolution of Anopheles species that were highly anthropophilic and effective *falciparum* vectors.[40,41,45]

If the age of Malaria's Eve is attributable to particular anthropological and/or epidemiological conditions, it may be that the sweep of SNP variation is due to more than merely a stochastic, demographic event. Strong selection for mutants that were favored under the novel conditions could have fixed alleles at various loci in close proximity to the selected locus. Initially, we argued that since the dearth of SNP variation is genome wide—being observed in at least 5 of the 14 *P. falciparum* chromosomes—that a selective sweep scenario seems unlikely explanation because it would require at least 5 independent events.[20] Of course, a single selective event could explain a genome wide sweep if it were the case the *P. falciparum* is asexual. Interestingly, while the parasite does have an obligate sexual phase in its life cycle, there are some indications that it has a largely clonal population structure, but this remains an issue of yet another contentious debate.[46-50]

If the P. falciparum genome is relatively youthful when we consider the level of SNP variation among largely neutral genes, then there is an apparent contradiction in the abundance of replacement changes observed in loci encoding antigenic determinants and drug resistance factors. It must first be noted that the polymorphisms in these antigenic genes, whether or not they are of ancient origin, do not contradict the recent origin of *P. falciparum* current world populations. Ancient polymorphisms at certain loci under strong balancing (diversifying) selection, can be maintained through a severe constriction in population numbers, or even through a number of generations with small populations that would lead to the virtual complete elimination of neutral allelic polymorphisms. For example, although the mitochondrial lineage of modern humans is only 100,000-200,000 years old, natural selection has maintained extensive polymorphisms among human *Hla* genes, some of which predate the split between humans and chimpanzees.[17,51] The *P. falciparum* antigenic genes are under strong diversifying selection for evasion of human immune response, and so they too could be maintained even through a demographic bottleneck.[18,52,53]

However, we have argued that it is not the age of selected genomic regions that sets them apart from the balance of the genome; rather, it is the rate at which these genes have mutated that makes them so unique. A notable feature of nearly every *P. falciparum* surface protein gene is the presence of repeat regions that encode short iterative amino acid sequences.[54,55] These antigenic repeat regions are highly polymorphic, yet are also known in many instances to be under immune selection. This presents a novel situation in molecular evolution since these loci behave as one would expect satellite DNA to behave with respect to the rapid mutation process and the generation of variable-length sequences, although the repeat portions encode part of the functional protein and so are subject to selection pressure. As with the DNA repeat regions that make up micro- and minisatellite loci in various plant and animal species, most of the variation within these repeats originates by a slipped-strand process that yields multiplication and/or deletion of the repeated units. This process leads to rapid differentiation of alleles, wherein an individual mutational event can change several nucleotides at once, with greater impact on sequence divergence than the typical single-nucleotide mutation process. Moreover,

these mutations occur at rates that are orders of magnitude greater than that of single nucleotide substitutions.[19,56] Based on the lack of silent site differentiation among dimorphic antigenic determinants and the occurrence of ancestral states for these divergent forms in the chimpanzee parasite, we have argued that the genes encoding antigenic determinants are evolving at extremely high rates due to the occurrence of repetitive regions.[31]

Not all repetitive regions in the P. falciparum genome are found in genes encoding antigenic determinants. Su and Wellems first described a large, highly polymorphic series (> 900) of microsatellite (Msat) markers which are distributed across the genome of P. falciparum.[35] Because of their abundance in the genome and high variability, microsatellite loci are ideally suited markers for estimating population parameters in natural populations. Following the genetic cross and genotyping of first generation progeny, 13 of 901 (1.4%) loci were found to contain mutations not evident in either parent. This suggests that these loci, as in many species of plants and animals, may have rather high mutation rates. Indeed, Su and Wellems' data would suggest that mutation rates of 10^{-1} may be reasonable (13 of 900 loci acquired mutation following a single generation, assuming mutation rate is equal across all loci).

Levels of polymorphisms in microsatellite loci are therefore extremely high relative to SNP's and since this variation is largely neutral,[22] Msat's should serve as genetic markers for resolving finer-scale chronologies of events following the origin of Malaria's Eve. In short, the virtual absence of SNP's means that they can provide no meaningful interpretation of recent events, but the much faster-evolving Msat loci will provide glimpses of post-Malarial Eve events.

Msat's have been used to infer a selective sweep in *P. falciparum* that is putatively associated with rise of chloroquine resistance sometime within the last half-century. Chloroquine is an important anti-malarial drug for which efficacy has waned in recent decades due to proliferation of resistant parasite strains.[57] Parasite resistance to chloroquine shows strong correlation to several point mutations in transmembrane vacuole protein (*PfCRT*) located on chromosome 7.[58] In a survey of Msat polymorphism within and between chloroquine-sensitive (CQS) and chloroquine-resistant (CQR) phenotypes, Wootton et al reported low levels of polymorphism among CQR strains in areas flanking the *PfCRT* as compared to distal regions of chromosome 7 and to the remainder of the genome.[59] Among the CQS strains, levels of Msat polymorphism are evenly distributed across the region, indicating that the sweep of Msat variability is highly correlated with chloroquine resistance. Based on the extent of linkage disequilibrium decay, the authors estimated that this sweep occurred with the last 20-80 generations. This corresponds roughly to between 6-30 years ago for the occurrence of the sweep, which is compatible with epidemiological studies of the emergence of global chloroquine resistance. This observation underscores both the remarkable adaptability of the parasite and also emphasizes the profound impact that human interventions can have on levels of polymorphism among its populations.

Summary

Studies of genomic polymorphism among *P. falciparum* populations have revealed that this parasite is surprisingly homogenous for most genetic loci. Past events in the evolution of the species have lead to vast reductions in genetic variation due to either demographic or selective sweeps, starting with the origin of Malaria's Eve perhaps as recently as 3,000 year ago. Subsequently, selective pressures such as those imposed by use of anti-malarial drugs, including chloroquine, may have further diminished variation over large regions of the genome. Nonetheless, it is clear that P. falciparum retains extraordinary ability to persist due in part to genomic novelties, such as repetitive antigen genes, which allow for rapid proliferation of genetic diversity despite demographic constraints. Understanding the precise mechanisms by which this adaptive potential is maintained and generated will be vital step in developing sustainable strategies to reduce malaria transmission and lessen the burden on human health.

References

1. Trigg P, Kondrachine A. The current global malaria situation. In: Sherman IW, ed. Malaria: Parasite Biology, Pathogenesis, and Protection. Washington DC: 1998:11-22.
2. Price RN. Artemisinin drugs: Novel antimalarial agents. Expert Opin Investig Drugs Aug 2000; 9(8):1815-1827.
3. Macreadie I, Ginsburg H, Sirawaraporn W et al. Antimalarial drug development and new targets. Parasitol Today Oct 2000; 16(10):438-444.
4. Guerin PJ, Olliaro P, Nosten F et al. Malaria: Current status of control, diagnosis, treatment, and a proposed agenda for research and development. Lancet Infect Dis Sep 2002; 2(9):564-573.
5. Moorthy V, Hill AV. Malaria vaccines. Br Med Bull 2002; 62:59-72.
6. Plebanski M, Proudfoot O, Pouniotis D et al. Immunogenetics and the design of Plasmodium falciparum vaccines for use in malaria-endemic populations. J Clin Invest Aug 2002; 110(3):295-301.
7. Ito J, Ghosh A, Moreira LA et al. Transgenic anopheline mosquitoes impaired in transmission of a malaria parasite. Nature May 23 2002; 417(6887):452-455.
8. Atkinson PW, Michel K. What's buzzing? Mosquito genomics and transgenic mosquitoes. Genesis Jan 2002; 32(1):42-48.
9. Walliker D, Carter R, Morgan S. Genetic recombination in malaria parasites. Nature 1971; 232(5312):561-562.
10. Walliker D. Genetic variation in malaria parasites. Br Med Bull 1982; 38(2):123-128.
11. Anders RF. Multiple cross-reactivities amongst antigens of Plasmodium falciparum impair the development of protective immunity against malaria. Parasite Immunol 1986; 8(6):529-539.
12. Ayala FJ, Escalante AA, Rich SM. Evolution of Plasmodium and the recent origin of the world populations of Plasmodium falciparum. Parassitologia 1999; 41(1-3):55-68.
13. Ayala F, Escalante A, Lal A et al. Evolutionary relationships of human malarias. In: Sherman IW, ed. Malaria: Parasite Biology, Pathogenesis, and Protection. Washington DC: American Society of Microbiology, 1998:285-300.
14. Ruvolo M. Molecular phylogeny of the hominoids: Inferences from multiple independent DNA sequence data sets. Mol Biol Evol Mar 1997; 14(3):248-265.
15. Hughes AL. Coevolution of immunogenic proteins of Plasmodium falciparum and the host's immune system. In: Takahata N, Clark AG, eds. Mechanisms of Molecular Evolution. Sunderland, MA: Sinauer Assoc., 1993:109-127.
16. Hughes MK, Hughes AL. Natural selection on Plasmodium surface proteins. Mol Biochem Parasitol 1995; 71(1):99-113.
17. Ayala FJ, Escalante A, O'Huigin C et al. Molecular genetics of speciation and human origins. Proc Natl Acad Sci USA 1994; 91(15):6787-6794.
18. Escalante AA, Lal AA, Ayala FJ. Genetic polymorphism and natural selection in the malaria parasite plasmodium falciparum. Genetics 1998; 149(1):189-202.
19. Rich SM, Ferreira MU, Ayala FJ. The origin of antigenic diversity in Plasmodium falciparum. Parasitol Today 2000; 16(9):390-396.
20. Rich SM, Licht MC, Hudson RR et al. Malaria's eve: Evidence of a recent population bottleneck throughout the world populations of Plasmodium falciparum. Proc Natl Acad Sci USA 1998; 95(8):4425-4430.
21. Gardner MJ, Hall N, Fung E et al. Genome sequence of the human malaria parasite Plasmodium falciparum. Nature Oct 3 2002; 419(6906):498-511.
22. Volkman SK, Barry AE, Lyons EJ et al. Recent origin of Plasmodium falciparum from a single progenitor. Science Jul 20 2001; 293(5529):482-484.
23. Hartl DL, Volkman SK, Nielsen KM et al. The paradoxical population genetics of Plasmodium falciparum. Trends Parasitol Jun 2002; 18(6):266-272.
24. Conway DJ, Fanello C, Lloyd JM et al. Origin of Plasmodium falciparum malaria is traced by mitochondrial DNA. Mol Biochem Parasitol 2000; 111(1):163-171.
25. Hughes AL, Verra F. Very large long-term effective population size in the virulent human malaria parasite Plasmodium falciparum. Proc R Soc Lond B Biol Sci Sep 7 2001; 268(1478):1855-1860.
26. Mu J, Duan J, Makova KD et al. Chromosome-wide SNPs reveal an ancient origin for Plasmodium falciparum. Nature Jul 18 2002; 418(6895):323-326.

27. Conway DJ, Baum J. In the blood—the remarkable ancestry of Plasmodium falciparum. Trends Parasitol Aug 2002; 18(8):351-355.
28. Arnot DE. Possible mechanisms for the maintenance of polymorphisms in Plasmodium populations. Acta Leiden 1991; 60(1):29-35.
29. Saul A, Battistutta D. Codon usage in Plasmodium falciparum. Mol Biochem Parasitol 1988; 27(1):35-42.
30. Saul A. Circumsporozoite polymorphisms, silent mutations and the evolution of Plasmodium falciparum. Parasitol Tod 1999; 15(1):38-39.
31. Rich SM, Ayala FJ. Population structure and recent evolution of Plasmodium falciparum. Proc Natl Acad Sci USA 2000; 97(13):6994-7001.
32. Rich SM, Ayala FJ. Reply to Saul. Parasitol Tod 1999; 15(1):39-40.
33. Anderson TJ, Su XZ, Roddam A et al. Complex mutations in a high proportion of microsatellite loci from the protozoan parasite Plasmodium falciparum. Mol Ecol 2000; 9(10):1599-1608.
34. Anderson TJ, Su XZ, Bockarie M et al. Twelve microsatellite markers for characterization of Plasmodium falciparum from finger-prick blood samples. Parasitology 1999; 119(Pt 2):113-125.
35. Su X, Wellems TE. Toward a high-resolution Plasmodium falciparum linkage map: Polymorphic markers from hundreds of simple sequence repeats. Genomics 1996; 33(3):430-444.
36. Forsdyke D. Selective pressures that decrease synonymous mutations in Plasmodium falciparum. Trends Parasitol Sep 2002; 18(9):411.
37. Sherman IW. A brief history of malaria and the discovery of the parasite's life cycle. In: Sherman IW, ed. Malaria: Parasite Biology, Pathogenesis, and Protection. Washington DC: American Society of Microbiology, 1998:3-10.
38. Tish KN, Pillans PI. Recrudescence of Plasmodium falciparum malaria contracted in Lombok, Indonesia after quinine/doxycycline and mefloquine: Case report. N Z Med J 1997; 110(1047):255-256.
39. Coluzzi M. Malaria and the Afrotropical ecosystems: Impact of man-made environmental changes. Parassitologia 1994; 36(1-2):223-227.
40. Coluzzi M. The clay feet of the malaria giant and its African roots: Hypotheses and inferences about origin, spread and control of Plasmodium falciparum. Parassitologia 1999; 41(1-3):277-283.
41. Coluzzi M. Evoluzione biologica & i grandi problemi della biologia: Accademia dei lincei. 1997:263-285.
42. Livingston FB. Anthropological Implications of sickle cell gene distribution in West Africa. Am Anthropol. 1958; 60:533-560.
43. Weisenfeld SL. Sickle-cell trait in human biological and cultural evolution. Development of agriculture causing increased malaria is bound to gene-pool changes causing malaria reduction. Science 1967; 157:1134-1140.
44. de Zulueta J. Malaria and ecosystems: From prehistory to posteradication. Parassitologia 1994; 36(1-2):7-15.
45. De Zulueta J, Blazquez J, Maruto JF. Entomological aspects of receptivity to malaria in the region of Navalmoral of Mata. Rev Sanid Hig Publica (Madr) 1973; 47(10):853-870.
46. Rich SM, Ayala FJ. The recent origin of allelic variation in antigenic determinants of Plasmodium falciparum. Genetics 1998; 150:515-517.
47. Rich SM, Hudson RR, Ayala FJ. Plasmodium falciparum antigenic diversity: Evidence of clonal population structure. Proc Natl Acad Sci USA 1997; 94(24):13040-13045.
48. Babiker HA, Ranford-Cartwright LC, Currie D et al. Random mating in a natural population of the malaria parasite Plasmodium falciparum. Parasitology 1994; 109(Pt 4):413-421.
49. Anderson TJ, Haubold B, Williams JT et al. Microsatellite markers reveal a spectrum of population structures in the malaria parasite Plasmodium falciparum. Mol Biol Evol 2000; 17(10):1467-1482.
50. Ferreira MU, Ribeiro WL, Tonon AP et al. Sequence diversity and evolution of the malaria vaccine candidate merozoite surface protein-1 (MSP-1) of Plasmodium falciparum. Gene Jan 30 2003; 304(1-2):65-75.
51. Ayala FJ. Adam, Eve, and other ancestors: A story of human origins told by genes. Pubbl Stn Zool Napoli II 1995; 17(2):303-313.
52. McCutchan TF, Waters AP. Mutations with multiple independent origins in surface antigens mark the targets of biological selective pressure. Immunol Lett 1990; 25(1-3):23-26.

53. Miller LH, Roberts T, Shahabuddin M et al. Analysis of sequence diversity in the Plasmodium falciparum merozoite surface protein-1 (MSP-1). Mol Biochem Parasitol 1993; 59(1):1-14.
54. Anders RF, Coppel RL, Brown GV et al. Antigens with repeated amino acid sequences from the asexual blood stages of Plasmodium falciparum. Prog Allergy 1988; 41:148-172.
55. Dame JB, Williams JL, McCutchan TF et al. Structure of the gene encoding the immunodominant surface antigen on the sporozoite of the human malaria parasite Plasmodium falciparum. Science 1984; 225(4662):593-599.
56. Hancock JM. Microsatellites and other simple sequences: Genomic context and mutational mechanisms. In: Goldstein DB, Schlötterer C, eds. Microsatellites, Evolution and Applications. Oxford: Oxford University Press, 1999:1-9.
57. Bloland PB. Drug resistance in malaria. Geneva: World Health Organization; 2001.
58. Fidock AD, Nomura T, Talley KA et al. Mutations in the P. falciparum digestive vacuole transmembrane protein PfCRT and evidence for their role in chloroquine resistance. Mol Cell 2000; 6(4):861-871.
59. Wootton JC, Feng X, Ferdig MT et al. Genetic diversity and chloroquine selective sweeps in Plasmodium falciparum. Nature Jul 18 2002; 418(6895):320-323.

CHAPTER 9

Selective Sweeps in Structured Populations—Empirical Evidence and Theoretical Studies

Thomas Wiehe, Karl Schmid and Wolfgang Stephan

Introduction

Properties of most selection models have been explored for panmictic populations, but little work has been done for substructured populations. In a substructured population, background selection against deleterious mutations has been shown to increase F_{ST}, a relative measure of differentiation between subpopulations, in chromosomal regions of low recombination because the effective size of local demes is reduced relative to that of high-recombination regions.[1,2] On the other hand, directional selection and genetic hitchhiking associated with the fixation of advantageous alleles may lead to greater homogeneity among populations if the selected allele causing the hitchhiking event in one deme migrates to other demes and causes a hitchhiking event in these demes as well (see Fig. 1). This scenario, however, is expected only in regions of zero or extremely low recombination. For larger recombination rates, different neutral variants may become linked to the selected allele in the population in which the advantageous mutation arose. In this situation, limited migration of the selected allele may lead to increased differentiation at linked neutral loci between populations.[3] Yet another case is that of local selection, in which the selected allele causing the hitchhiking event is locally adapted. Here, hitchhiking events are assumed to be restricted to single demes or parts of the species range and, as a consequence, may cause substantial genetic differences between populations.[2,4,5]

To test these hypotheses about the migration behavior of selected genes and to compare it to that of neutral genes, we utilize the facts (1) that in some well-characterized sexually reproducing species, such as Drosophila,[6] recombination rates vary drastically along chromosomes, and (2) that the evolutionary dynamics of genes in regions of high recombination rates do not generally deviate from a (nearly) neutral model, whereas genes in regions of reduced recombination may exhibit footprints of natural selection due to linkage to selected loci. The high-recombination loci may therefore serve as neutral markers in the analysis of past selective or demographic events. As a sexually reproducing species showing extensive population structure we discuss *Drosophila ananassae*, and also mention some human examples. Finally we use the highly selfing *Arabidopsis thaliana* and its close outcrossing relative *A. lyrata* to illustrate the situation in plants.

Selective Sweep, edited by Dmitry Nurminsky. ©2005 Eurekah.com and Kluwer Academic/Plenum Publishers.

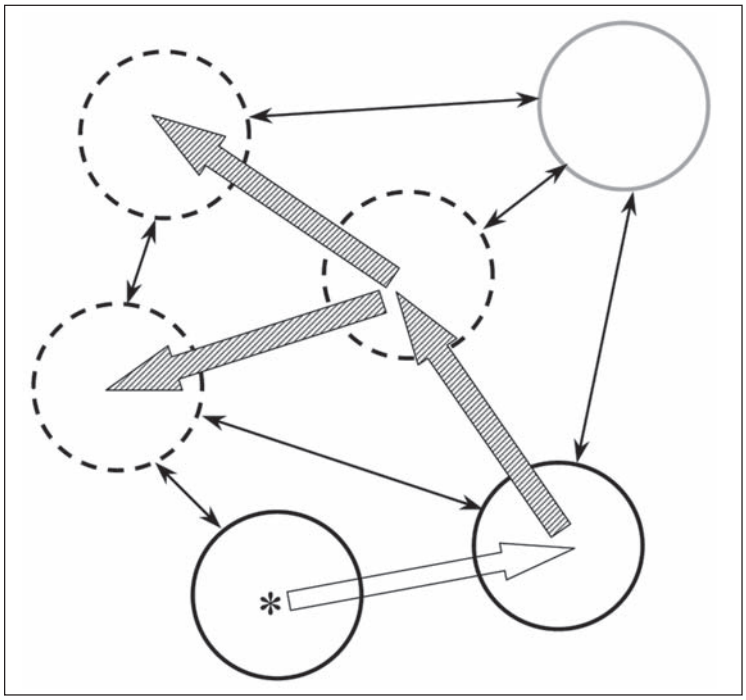

Figure 1. Symmetric island model. A sweep allele, which originated in a single subpopulation (marked by a star), is exported by migrants to other subpopulations. There are three types of subpopulations: 1) those in which the original neutral marker allele, say *A*, is linked to the sweep allele (solid circles), 2) those in which the neutral allele *a* is linked to the sweep allele (dashed circles) and 3) those in which no sweep has taken place yet (grey circle). For type-2 populations to be present, the sweep allele has to be decoupled from *A* by recombination. Migration between some of the subpopulations is indicated by double-head arrows.

Experimental Evidence

Data from Drosophila Ananassae

Drosophila ananassae is an ideal organism for our purposes because it exists in highly structured populations and has been used extensively in genetic analysis (reviewed by Tobari[7]). Other Drosophila species for which detailed genetic maps are available (*D. melanogaster* and *D. simulans*) have either low levels of population structure throughout their distribution or are geographically restricted and show little evidence of population structure. *D. ananassae* is the most abundant Drosophila species in much of the tropical and subtropical regions of the world and has even been observed in the milder American climatic regions.[8] Its geographic center is thought to be in Southeast Asia,[7] and it has most likely colonized much of the world very recently, invading a variety of climatic zones. It currently exists in many semi-isolated populations in the geographic regions where it has been studied.[9-13] Population structure is evident along clines in India,[14] and is particularly strong among the island populations in the south Pacific Ocean.[9,13] Detailed cytological and genetic maps based on polytene chromosomes and visible mutants have been constructed for *D. ananassae*.[7,15] These maps have demonstrated a centromeric effect (reduced recombination) on the X chromosome providing a means for us to compare DNA sequence variation of X-linked genes in regions of reduced recombination and compare them with X-linked genes in areas with intermediate or high rates of recombination.

Population Structure of *D. Ananassae* Inferred from a High-Recombination Locus

We measured variation in a 1.8-kb segment of *Om(1D)* located in a region of normal to high recombination on the X chromosome.[16] As explained before, DNA sequence variation of genes in such regions is likely to evolve according to a neutral model because recombination is expected to break up linkage disequilibrium with selected loci. In fact, patterns of variation at *Om(1D)* were consistent with a neutral equilibrium model. We thus used this gene to measure gene flow between two populations from southern, subtropical areas [Hyderabad (India) and Sri Lanka] and two from temperate zones in the north (Nepal and Myanmar) along a longitudinal transect on the Indian subcontinent. Average nucleotide diversity at *Om(1D)* is around 0.01 within these populations, similar to estimates obtained from RFLP analysis for *Om(1D)*[11] and *forked*[10] which maps to the same polytene band as *Om(1D).* Tests for detecting genetic differentiation between pairs of populations revealed that all four populations are genetically distinct at the neutral *Om(1D)* locus and show a pattern of isolation-by-distance.[17]

Does Natural Selection Affect DNA Sequence Variation at Low-Recombination Loci in *D. Ananassae*?

To test for evidence of natural selection on the X chromosome, we followed the general method outlined in the Introduction and assayed two loci in regions of the X chromosome with low rates of recombination within and between the same four populations of *D. ananassae* analyzed for *Om(1D)*. The first is a 3.6-kb region encompassing the *vermilion* (*v*) gene[16] and the second is a 5.7-kb fragment of the *furrowed* (*fw*) gene region.[18] *v* is located in a region of low recombination in the first clearly visible polytene band to the left of the centromere on the X chromosome. *fw* is in a region of presumably even lower rate of recombination on the right arm, as it maps to the transition zone at the base of the X between β-heterochromatin and euchromatin.[4] Average nucleotide diversity within populations is 20-50 times lower in *v* and *fw* than in regions with normal to high rates of recombination. Levels of divergence between *D. ananassae* and a sibling species, *D. pallidosa*, are similar between *v* and *Om(1D).* Divergence at *fw* was lower than at *Om(1D)*, but still at least 10-25 fold higher than nucleotide diversity within populations. For both *v* and *fw*, a constant-rate, neutral model of molecular evolution is rejected[19] providing evidence that reduced variation is due to natural selection. The much lower level of variation at the low-recombination loci indicates that natural selection has a strong effect on levels of X chromosome variation in *D. ananassae*.

We also compared population structure at *fw* and *v* with that at *Om(1D)*. In contrast to *Om(1D)*, *v* and *fw* show no differentiation between the northern two populations or between the southern two populations but strong differentiation between northern and southern populations. This pattern of no differentiation within a geographic region, but high differentiation between geographic regions is due to nearly fixed differences between the northern and southern populations for about half of the polymorphisms observed.[16,18] Together these results raise the intriguing possibility that patterns of molecular variation on the X chromosome reflect the effects of natural selection in different geographic regions.

Distinguishing Between Alternative Models of Natural Selection: Selective Sweep Versus Background Selection

Both a selective sweep model and a background selection model may explain the general observation of reduced variation in regions of low recombination (see above), and the strong differentiation between northern and southern populations at low recombination genes.[4,20,21] The selective sweep model assumes differentiation in regions of low recombination is due to the locally favored substitutions.[4,5] In contrast, the background selection model assumes that the differentiation is caused by the continual removal of deleterious alleles in regions of low recombination which results in lower effective population sizes and thus in a lower migration rate for the low-recombination locus.[1]

We developed a statistical test to distinguish between the selective sweep and the background selection hypotheses in structured populations,[16,18] described below ("Theoretical Studies" section). Applying this test to the *v* and *fw* data indicates that, in addition to deleterious mutations that undoubtedly occur (reviewed in Keightley and EyreWalker[22]), one has to postulate the occasional occurrence of advantageous mutations. Indeed, the observed pattern of differentiation at *fw* is consistent with recent selective sweeps that homogenized single nucleotide polymorphism frequencies within the northern as well within the southern populations. At the *v* locus, only the data from the northern populations rejected the background selection model.

Whether a single sweep model[3] can explain the data at *fw* or independent sweeps must be invoked[4] is difficult to decide. A single sweep model seems to be the most parsimonious explanation. But for a single sweep model to explain the data, at least two haplotypes of the *fw* locus must be linked to the advantageous allele (see section "Sweeps and Population Structure" and Fig. 1). This would require that the levels of nucleotide variation before the hitchhiking event were sufficiently high and that the sweep took long enough to spread through the northern and southern populations for different haplotypes to become associated with the advantageous mutation via recombination. The occurrence of nearly fixed differences between populations can be explained by both a single and multiple sweep model. To determine if independent, local selective sweeps have occurred in the northern and southern populations, a larger number of polymorphic sites in these low recombination regions is required to obtain accurate estimates of the frequency spectrum of polymorphisms.

Evidence for adaptation of *D. ananassae* populations to local environmental conditions in India has also been documented at the phenotypic level. For ectothermic organisms such as Drosophila, temperature is an important environmental factor. In particular, *D. ananassae* is known to be stenotherm and cold sensitive, presumably due to its tropical origin.[23] At low-altitude localities along the Indian subcontinent, average yearly temperature is relatively constant ranging from 24ºC to 27ºC, but seasonal variation increases dramatically with latitude. In *D. ananassae*, several morphometric traits such as wing length, thorax length, and ovariole number show latitudinal clines which may be caused by temperature adaptation.[24] Similar latitudinal clines were observed for desiccation tolerance and starvation tolerance.[25,26]

Humans

Resistance to Malaria

In humans, malaria has been a strong selective agent, leading to high-frequency advantageous polymorphisms at several genes (e.g., Duffy blood group locus, α-globin, β-globin, G6PD). As an example of local adaptation (resistance to malaria), we discuss here the Duffy factor (FY). This gene is different from the other known genes involved in malaria resistance. The mutations producing the well-known sickle allele in the β-globin gene and the G6PD A-allele are deleterious in the absence of malarial selection.[27] That means these mutations are found as balanced polymorphisms within local populations occurring at frequencies that are correlated with the local incidence of malaria. In contrast, the FY*O mutation of the Duffy gene is (nearly) fixed in sub-Saharan Africa, but absent or in low frequency in other geographic regions. This suggests that this FY mutation that confers resistance to malaria caused by *Plasmodium vivax* infection was a target of local directional selection pressures (local adaptation), whereas the alleles of the other genes conferring resistance to malarial infections due to *P. falciparum* (listed above) were more likely subject to balancing selection.

Hamblin and Di Rienzo[28,29] set out to test this hypothesis for the Duffy locus. The FY*O mutation in sub-Saharan Africans represents a best-case scenario for detecting the effects of directional selection on patterns of sequence variation: the phenotype is clear and well

understood, the precise nucleotide location of the responsible mutation is known, and the population is more variable and closer to equilibrium than non-African populations. Except for the fact that the FY locus is in a chromosomal region of normal recombination, the authors used a similar approach as outlined above for *D. ananassae* by comparing DNA sequence variation at the Duffy gene in several populations to patterns of variation at putatively neutral marker genes. This study produced the following results. In the sample of the sub-Saharan Hausa population for which the allele FY*O was fixed, the level of DNA sequence variation was reduced around the site of the FY*O mutation relative to the flanking regions (including up to 16 kb on both sides). Furthermore, genetic differentiation between the Hausa population and non-African populations (measured by F_{ST}) near the position of the FY*O mutation was significantly higher than that observed for the neutral marker genes and decreased at greater distances (> 10 kb) from the site of mutation. These results are consistent with the simple model of genetic hitchhiking caused by local directional selection. Using this model, the authors estimated the time of fixation of the FY*O mutation as approximately 0.16 N_e generations and its selection coefficient as 0.002 or larger.

Additional Examples of Locally Adapted Traits in Humans

Other examples of locally adapted traits in humans include skin pigmentation and lactose tolerance. However, to our knowledge, specific analyses documenting the occurrence of local selection have not been performed in these examples. In the first case, a survey of the melanocortin 1 receptor locus in a worldwide sample found high nucleotide diversity relative to the average nucleotide diversity in human populations.[30] Variation occurred mostly at nonsynonymous sites. Certain alleles were very frequent in some populations but nearly absent in others. The Arg163Gln variant (absent in the Africans studied) seems to have risen to high frequency (of about 70% in East and Southeast Asians) only very recently, possibly as a result of local adaptation.

Lactase persistence, the trait in which intestinal lactase activity persists at childhood levels into adulthood, varies in frequency in different human populations, being most frequent in northern Europeans and certain African and Arabian nomadic tribes, who have a history of drinking fresh milk. It has been shown that the element responsible for the lactase persistence/nonpersistence polymorphism is cis-acting to the lactase gene and that lactase persistence is associated with the most common 70-kb lactase haplotype. Directional selection has been postulated to explain the association of this haplotype with lactase persistence in northern Europeans.[31]

Arabidopsis Thaliana *and Close Relatives*

Arabidopsis thaliana has recently become a model for plant population genetics because of its important role in plant functional genomics.[32] Many traits show a high degree of naturally occurring, inheritable and potentially adaptive phenotypic variation among accessions that is very useful in mapping the genes underlying these traits.[33] Two characteristics of *A. thaliana* need to be considered in studies of local selective sweeps. First, this species is almost completely selfing (more than 99%),[34] which leads to low effective recombination rates and a reduced effective population size. Both selective sweeps and background selection are then expected to reduce genome-wide levels of nucleotide diversity considerably.[35,36] Due to reduced effective population size, natural selection is expected to be less effective in removing slightly deleterious mutations from the population. Another consequence of selfing is that combinations of favorable alleles may not be broken up easily by recombination, which favors local adaptation to habitat-specific environments.

Second, *A. thaliana* is an annual species that tends to live at disturbed sites with low competition to other species. Thus, one can expect that local populations consist of unstable

metapopulations founded from few, highly inbred individuals. Such a metapopulation structure is consistent with the results from several surveys using RFLP and AFLP markers. A RFLP survey of more than 100 individuals from ten local populations detected only very little variation within, but much variation between populations as indicated by high F_{ST} values.[37] There was no correlation between genetic and geographical distance of populations, and the individual haplotypes appear to have a worldwide distribution. This pattern was confirmed by two AFLP studies[38,39] although there was also evidence for some population structure resulting from isolation-by-distance. Coalescent simulations using these data fit better with a model of an exponentially growing metapopulation than a model of constant population size.[40] These data suggest that in *A. thaliana* there has been long-distance gene flow or a recent population expansion, both of which could have been associated with the spread of human agriculture.

Despite the apparent lack of population structure, there are a number of traits showing variation across a geographic cline. Such clines may result from local adaptation and the genes responsible for these traits may be targets of locally restricted selective sweeps. For example, flowering time follows a gradient from north to south across Europe. Late flowering accessions that need a vernalization treatment occur predominantly in Northern Europe whereas early flowering accessions have been mostly collected from Central and Eastern Europe.[41] This pattern can be interpreted as vernalization being advantageous in Northern latitudes because it allows plants to survive long winters. The molecular analysis of the *FRIGIDA* (*FRI*) locus demonstrated that late flowering is the ancestral condition and early flowering results from recessive loss-of-function alleles in *FRI* (or other flowering time loci). Johanson et al[41] concluded that there was a strong selective pressure in Central Europe for an early flowering phenotype. The *FRI* gene appears to have been the prime target of selection because most natural early flowering accessions contain a loss-of-function allele of *FRI* which appears to have the greatest phenotypic effect on flowering time and fewer deleterious pleiotropic effects compared to other flowering time genes.

More recently, Le Corre et al[42] have analyzed patterns of polymorphism at *FRI* in 25 accessions from France and England to test whether these patterns are consistent with the hypothesis of local selection for early flowering. They find high levels of amino acid polymorphism and low level of silent polymorphism. Tests of neutrality, such as Tajima's[43] test and the McDonald-Kreitman[44] test, reject a neutral evolution model at this locus. In addition, they find eight mutations that lead to a loss-of-function allele and all of these mutations are associated with an early flowering phenotype. A haplotype network of the sequences suggests a recent origin of the loss-of-function alleles (and, in fact, of all earliness phenotypes) suggesting that selection for earliness has been recent, possibly after the last glaciation.

One intriguing result of this study, however, is that although the data look compatible with a local selective sweep at a single locus, the pattern of variation may not be caused by the fixation of a single advantageous allele. Since the advantageous phenotype results from loss-of-function mutations and there are many possibilities to knock out a gene, any given advantageous knock-out allele may not have enough time to be swept to fixation. New loss-of-function mutations are constantly being generated at a fairly high rate, which leads to a large number of different haplotypes such as observed by Le Corre et al.[42] Under this scenario one also expects an excess of rare polymorphisms leading to a significant result of Tajima's test but for a different reason than under a selective sweep model.

In summary, despite some evidence for local phenotypic adaptation in *A. thaliana* mentioned above (see also the review by Pigliucci[45]), this species does not appear to be a good model system for studying the population genetics of local adaptation because of its selfing nature and metapopulation structure. Both factors are not favorable for F_{ST}-based tests. *A. thaliana* may not have enough population structure to exclude the case where several advantageous mutations are competing for fixation at the same time and lead to a 'traffic' situation if

genomic regions that hitchhike with different sweeps overlap with each other.[46] Currently, there is no suitable statistical test of the traffic model available.

Arabidopsis lyrata ssp. *petraea*, a self-incompatible perennial, may be more suited for studying local selection than *A. thaliana* because it occurs in spatially restricted and isolated populations throughout Europe. Gene flow among populations may be low and increase the chance of local adaptation, which may make models of "single sweeps" applicable. Currently available data are consistent with such a scenario. For example, a survey of 35 microsatellite loci in a single population from Central Europe revealed a high degree of genetic variation in comparison to *A. thaliana*: Heterozygosity was about 40 times higher in *A. petraea* (0.041) than in *A. thaliana*[47] (0.001). Simulations revealed that the distribution of allele frequencies in this population are similar to the expectation under a mutation-drift equilibrium model, which suggests that the population is old and stable. A survey of nucleotide diversity at the Myrosinase locus in populations of the *lyrata* subspecies of *Arabidopsis lyrata*, which occurs in North America, also suggests a distinct population structure because F_{ST} values are in the range of 0.45 for this subspecies.[48]

Theoretical Studies

Estimating Population Subdivision

Statistical measures of genetic differentiation of subpopulations employ allele frequencies of polymorphic loci in samples taken from different localities[49] (for recent reviews, see chapters 9 and 10 in Balding et al.[50] One of the standard measures is Wright's[17,51] F_{ST}, which is a reformulation of the traditional inbreeding coefficient F in terms of a ratio of allele frequencies in demes, or subpopulations, and the total population. It measures the deviation from Hardy-Weinberg equilibrium of the total population. More recently, instead of allele frequency based statistics, sequence based statistics have been used to measure genetic differentiation.[52] These statistics are more appropriate for DNA sequence data collected today. The simulation results by Hudson[52] show that the potential to detect local differentiation with sequence based statistical tests is sensitive to recombination and mutation rates, with the power generally decreasing when mutation rates decrease. With respect to the recombination rate there is no uniformly valid trend among these tests, but most tests investigated show less power for lower recombination rates.

Given that genetic differentiation between subpopulations was detected, the next question concerns its cause. It may be a mere consequence of restricted migration. Wright's formula

$$F_{ST} = 1/(1+4N_e\, m),$$

where N_e is the effective population size and m the migration rate per generation, describes the migration-drift equilibrium for a symmetric island population. If the magnitude of N_e or m (or both) is unknown, how can this situation be distinguished from one in which natural selection causes genetic differentiation? Purely demographic causes should affect the entire genome more or less uniformly. On the other hand, the footprints of selection should be confined to a more or less restricted segment of the genome—depending on the recombination rate. The answer may come from a comparative approach. As mentioned above, loci in regions of high recombination rates may be used as neutral markers and measurement of genetic differentiation (e.g., F_{ST}) at these loci can be compared with those from rarely recombining genes. Discrepancies in the magnitude of F_{ST} should be indicative of nonneutral evolutionary mechanisms. Generally, one may collect data from many genomic segments, perhaps entire chromosomes, and plot an F_{ST} profile along a chromosome. Local adaptation, selective events which are confined to a particular subpopulation, should be detectable if such profiles show a

significant deviation from their mean value. This approach has been used by Hamblin et al[28] for analysis of the Duffy locus in humans (described in "Humans" section above).

Profiles of genetic variability which may be characteristic of a given subpopulation have first been explored by Charlesworth et al[2] both analytically and in computer simulations. Since the effects of background selection[20] as well as of genetic hitchhiking, discussed below, critically depend on the recombination rate, these selective mechanisms should induce a characteristic signature along the chromosome, which may be uncovered if the level of genetic differentiation can be measured and compared for many loci along a chromosome.

In any case, there is a need for specifically designed statistical tests to detect selection in substructured populations. For the case of background selection, an example of such a test is described below. In the case of selective sweeps, a test may be constructed which considers the pattern of variation as a function of the physical distance from a selected site. Such a test for the detection of selective sweeps in panmictic populations has recently been described by Kim and Stephan.[53] To our knowledge, a similar test for subdivided populations is not available. In contrast, modeling and analytical work to describe the effect of selective sweeps on F_{ST} in a subdivided population has begun[3] (see section "Sweeps and Population Structure").

Test of the Background Selection Model in a Substructured Population

For small samples, background selection generates genealogies that are approximately identical to those produced by a strict neutral model if the effective population size is adjusted such that the effects of recombination and background selection on the locus of interest are taken into account.[1] The slight distortions of the allele frequency spectrum produced by background selection[20,54] are neglected because these can only be observed in rather large samples (usually not used). The effect of background selection on neutral variation in a substructured population can thus be analyzed by simulating a neutral coalescent in an appropriate model of population structure. As a starting point, we used the symmetric finite island model (Crow,[55] Chapter 3.4) with background selection incorporated as the general framework of our simulations.[16,18]

We investigated several statistical properties of this test including critical values and achieved levels of significance.[18] We also investigated the influence of the underlying model of population structure on the statistical power. The power of this test is expected to increase when a stepping-stone model rather than a finite island model is incorporated. Such a model is more realistic for *D. ananassae* populations which tend to show a pattern of isolation by distance.[18]

Selective Sweep Modeling

The first explicit model of the effects of a selective sweep on linked polymorphism was formulated by Maynard Smith and Haigh[56] in a deterministic setting. Later, Ohta and Kimura[57] studied a similar model. They introduced diffusion processes and were able to model the effects of random drift. Stephan et al[58] obtained analytical results, based on diffusion theory, for the reduction of heterozygosity due to a single selective sweep and recurrent sweeps. However, it is very difficult to generalize these results and to take additional features such as demographic effects and population structure into account. The diffusion process can be mimicked by forward-in-time, whole population simulations. However, such computer simulations are extremely time- and space-demanding. Therefore, coalescent simulations, which run backwards in time and only simulate the genealogy of population samples rather than whole populations, became the favored approach to model selective sweeps. Coalescent based hitchhiking models were introduced by Kaplan et al.[59] Kim and Stephan[53] studied the effect of a selective sweep on a genome wide scale with the help of a newly designed simulation program based on the so-called ancestral recombination graph.[60] These hitchhiking models assume a panmictic population with random mating. A comprehensive review of the current knowledge in the theory of genetic hitchhiking has recently been published by Barton.[61]

Sweeps and Population Structure

An explicit model of selective sweeps in subdivided populations was developed by Slatkin and Wiehe.[3] They used results from the deterministic analysis by Maynard Smith and Haigh[56] and extended them to a selective sweep scenario in a symmetric finite island model; i.e., it is assumed that a fixed fraction m of each subpopulation is replaced by immigrants drawn randomly from finitely many other subpopulations. In their model they consider the case where, initially, a selective sweep occurs in a single subpopulation, chosen at random. Eventually, a selectively favored individual emigrates and triggers a sweep in another island by importing the allele which had been favorable in the original island. Critical for creating population subdivision is that the emigrant haplotype which carries the favorable mutation from one deme to the next is different from the one in which the favorable mutation was originally present. In a simple deterministic two-locus two-allele model, the probability for this event is given by

$$(1-q)\, p^{c/s},$$

where q is the frequency of one of the alleles, A, at the neutral locus in any deme before the selective sweep, p is the frequency of the sweep allele B at the time when it arises in the first deme, c the recombination rate between both loci and s the selective advantage of the favored allele. After the sweep, the frequency of the neutral allele A in deme d will be

$$q' = q+(1-q)p^{c/s},$$

if B was linked with A in deme d, and it will be

$$q'' = q(1-p^{c/s}),$$

if B was linked with the other neutral allele, a.

After a while, all demes will have experienced this selective sweep, each one targeted with some time delay with respect to the previous one. The authors show that population subdivision, measured in terms of F_{ST} at the linked neutral marker locus, can be transiently increased through differential association of one of the neutral alleles with the favored allele. Slatkin and Wiehe[3] calculated also the dynamical behavior of F_{ST} and found that, as a function of time, F_{ST} passes through a maximal value and eventually decays on a time scale of $1/m$ generations to its drift-migration equilibrium state. The numerical value of the peak, the decay rate and the equilibrium depend on the selection coefficient, the migration and recombination rates. Their analytical results for a symmetric two-deme model show that F_{ST} is maximal for intermediate recombination rates, but is close to zero for relatively high migration rates ($2Nm > 1$) and for very small as well as for high recombination rates (Fig. 2) (see also Fig. 3 in Slatkin and Wiehe[3]). The reason is that loci which are very closely linked to the sweep locus are unlikely to be decoupled while the selective allele is being fixed. Therefore, differentiation will be low because the selective sweep induces uniformity of the tightly linked neutral alleles among demes. In contrast, very distant loci are likely to be decoupled by recombination during the selective phase and their level of polymorphism remains unaffected. The result, insofar as F_{ST} is concerned, is the same: there is no genetic differentiation among demes.

The authors also considered the propagation of a selective sweep in a one dimensional stepping stone model. Again, population subdivision is measured in terms of F_{ST} at the neutral marker locus. After the advantageous mutation has swept through all population patches, there are again two types of patches: (1) those in which the favorable allele B was linked to neutral allele A, leading to an increase in the frequency of A in type 1 demes, and (2) those, in which B was originally linked to neutral allele a, leading to a decrease of the frequency of A in type 2 demes, according to the formulae above. Going from patch to patch, type 1 and type 2 demes alternate according to a geometric distribution with parameters q' and q'', respectively. Therefore, although a selective sweep may create genetic differentiation among demes, a (geographically)

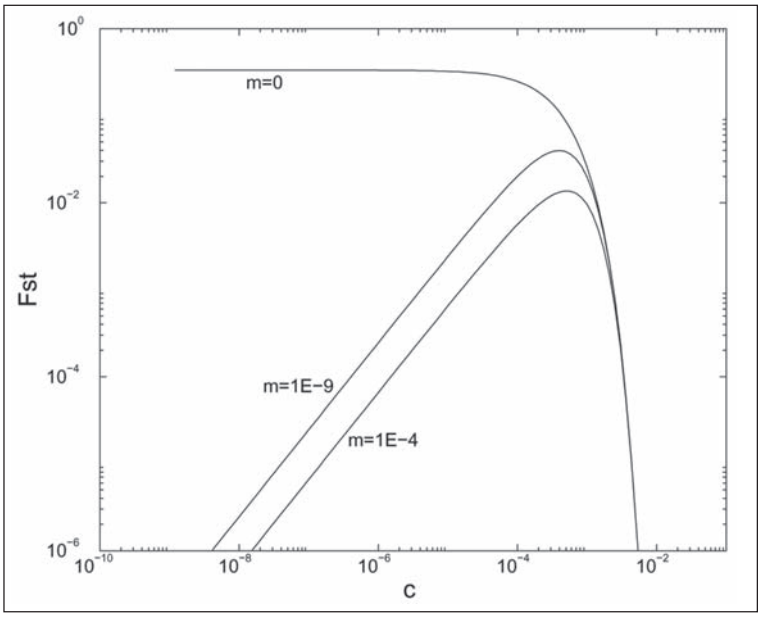

Figure 2. Symmetric island model with *d=2* subpopulations. Plot of F_{ST} versus the recombination rate *c* between the neutral marker locus and the sweep locus for various migration rates *m.* With some migration (*m>0*), population differentiation (F_{ST}) is maximal for some intermediate recombination rates. The location (not the height) of the maximum depends on the relative magnitudes of the selection coefficient and the recombination rate. In the absence of migration F_{ST} is a monotonically decreasing function of the recombination rate. The numerical values are obtained from eq. (20) in Slatkin and Wiehe.[3]

distant patch may be genetically more uniform to a given patch than a close one. Thus, a selective sweep may not induce isolation by distance (see Fig. 3).

Discussion

Our discussion of empirical evidence for local adaptation has shown that not every species may be suited for studying the effects of local adaptation on patterns of genetic variation. Due to the requirements and limitations of theoretical models of local selective sweeps, ideal species used to test these models should fulfill a set of requirements. First, they should be structured in well characterized and fairly well isolated subpopulations close to a mutation-drift equilibrium. Second, local populations should not result from recent population expansion because this would lead to founder effects. Third, populations should live in ecologically diverse habitats so that one can expect divergent adaptive evolution to happen. Fourth, the population structure should be stable. In particular, metapopulations with high extinction-recolonization dynamic may not persist long enough to show traces of selective sweeps within populations. Finally, the species should be closely related to a model organism for genome research so that the tools and resources developed for the model can be used to characterize the function of genes involved in local adaptation. Most of these requirements are fulfilled by *Drosophila ananassae* and (apparently) *Arabidopsis lyrata.*

In addition to the investigation of model organisms, the study of genetic variation involved in local adaptation is of particular interest in organisms that are of economical value. For example, populations of wild relatives of many crop species exhibit phenotypic variation in numerous traits in response to differences in local biotic (e.g., pathogenes) and abiotic (e.g.,

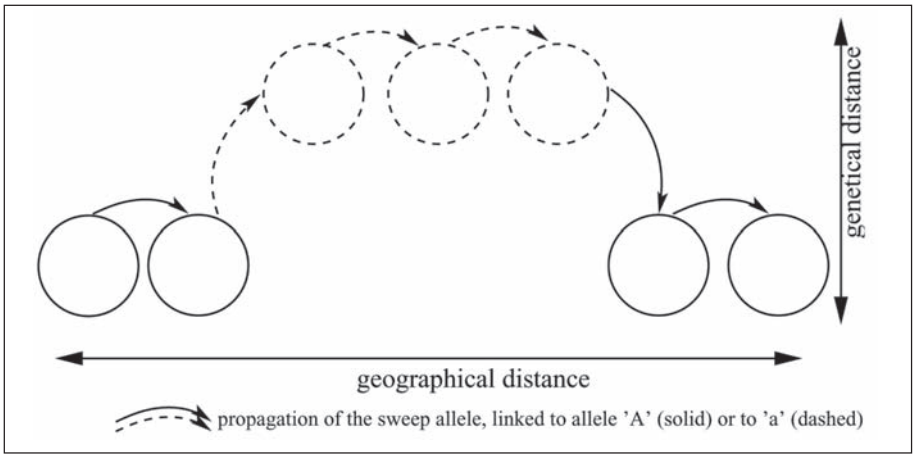

Figure 3. One dimensional stepping stone model. The propagation of the sweep allele by migration generates subpopulations of three types, as in Figure 1. Note, that geographical distance (x-axis) is not necessarily correlated with genetical distance (y-axis).

soil conditions) environments that may result from adaptive evolution. The genome-wide identification of genetic variation in such species using modern genomics tools, and the application of theoretical models of local sweeps in the analysis of this variation may prove to be useful as a heuristic tool for the mapping of genes involved in local adaptation. These genes could subsequently be introgressed into elite germplasm by molecular breeding approaches and contribute to the development of improved seeds for agriculture.

On the theoretical side, the biggest problem we are facing is that many statistical and mathematical methods are not readily applicable to species with substructure. For instance, the available statistical tests for neutrality, such as the HKA,[19] Tajima's[43] and McDonald-Kreitman[44] tests, are designed for panmictic, not for structured populations. Furthermore, the dynamics of selective sweeps in structured populations have so far only been analyzed in very simple models.[3] More general models need to be investigated. When considering selective sweeps in structured populations there are two main scenarios: first, local sweeps which are eventually exported from the deme of their origin to other demes by migration ('hitchhiking in space') and, second, independent local sweeps which remain confined to their deme of origin ('local adaptation'). Of particular interest would be a method to distinguish both types of sweeps. Preliminary analysis indicates that the unimodal profile of F_{ST} when plotted as a function of the recombination rate (see Fig. 2) and observed under the hitchhiking-in-space model loses its mode and turns into a monotonically decreasing function under the local-adaptation model. This behavior may be used as a basis for a statistical test to distinguish both models in regions of normal recombination rates.

Acknowledgments

This work was supported by a grant (Förderkennzeichen 0312705A) from the German Ministry for Education and Research (BMBF) to T.W., and by a grant from the German Research Foundation (DFG, grant no. Ste325/4) to W.S.

References

1. Hudson R, Kaplan N. Deleterious background selection with recombination. Genetics 1995; 141:1605-1617.
2. Charlesworth B, Nordborg M, Charlesworth D. The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. Genet Res 1997; 70:155-174.
3. Slatkin M, Wiehe T. Genetic hitch-hiking in a subdivided population. Genet Res 1998; 71(2):155-160.
4. Stephan W, Mitchell SJ. Reduced levels of DNA polymorphism and fixed between-population differences in the centromeric region of Drosophila ananassae. Genetics Dec 1992; 132(4):1039-1045.
5. Begun D, Aquadro C. African and north American populations of Drosophila melanogaster are very different at the DNA level. Nature 1993; 365:548-550.
6. Lindsley DL, Sandler L. The genetic analysis of meiosis in female Drosophila melanogaster Philos Trans R Soc London B Biol Sci 1977; 277:295-312.
7. Tobari YN. Drosophila ananassae - Genetical and Biological Aspects. Tokyo: Japan Scientific Societies Press, 1993.
8. Dobzhansky T, Dreyfus A. Chromosomal aberrations in Brazilian Drosophila ananassae. Proc Natl Acad Sci USA 1943; 29:301-305.
9. Johnson FM. Isozyme polymorphisms in Drosophila ananassae: Genetic diversity among island populations in the South Pacific. Genetics 1971; 68:77-95.
10. Stephan W, Langley CH. Molecular genetic variation in the centromeric region of the X chromosome in three Drosophila ananassae populations. I. Contrasts between the vermilion and forked loci. Genetics Jan 1989; 121(1):89-99.
11. Stephan W. Molecular genetic variation in the centromeric region of the X chromosome in three Drosophila ananassae populations II. The Om(1D) locus. Mol Biol Evol Nov 1989; 6(6):624-635.
12. Lynch M, Crease TJ. The analysis of population survey data on DNA sequence variation. Mol Biol Evol Jul 1990; 7(4):377-394.
13. Tomimura Y, Matsuda M, Tobari YN. Polytene chromosome variations of Drosophila ananassae and its relatives. In: Tobari YN, ed. Drosophila ananassae - Genetical and Biological Aspects. Tokyo: Japan Scientific Societies Press, 1993:139-151.
14. Singh BN. Population genetics of inversion polymorphism in Drosophila ananassae. Indian J of Exp Biol 1998; 36:739-748.
15. Tobari YN, Goñi B, Tomimura Y et al. Chromosome. In: Tobari YN, ed. Drosophila ananassae - Genetical and Biological Aspects. Tokyo: Japan Scientific Societies Press, 1993:23-48.
16. Stephan W, Xing L, Kirby DA et al. A test of the background selection hypothesis based on nucleotide data from Drosophila ananassae. Proc Natl Acad Sci USA May 12 1998; 95(10):5649-5654.
17. Wright S. Isolation by distance. Genetics 1943; 28:114-138.
18. Chen Y, Marsh BJ, Stephan W. Joint effects of natural selection and recombination on gene flow between Drosophila ananassae populations. Genetics 2000; 155:1185-1194.
19. Hudson RR, Kreitman M, Aguadé M. A test of neutral molecular evolution based on nucleotide data. Genetics 1987; 116:153-159.
20. Charlesworth B, Morgan M, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. Genetics 1993; 134:1289-1303.
21. Aquadro C, Begun D, Kindahl E. Selection, recombination, and DNA polymorphism in Drosophila. In: Golding B, ed. NonNeutral Evolution. London: Chapman and Hall, 1994:46-55.
22. Keightley PD, EyreWalker A. Terumi Mukai and the riddle of deleterious mutation rates. Genetics 1999; 153:515-523.
23. Cohet Y, Vouidibio J, David JR. Thermal tolerance and geographic distribution: A comparison of cosmopolitan and tropical endemic Drosophila species. J Therm Biol 1980; 5:69-74.
24. Morin JP, Moreteau B, Pétavy G et al. Reaction norms of morphometrical traits in Drosophila: Adapative shape changes in a stenotherm circumtropical species. Evolution 1997; 51:1140-1148.
25. Das A, Mohanty S, Parida BB et al. Variation in tolerance to starvation in Indian natural populations of Drosophila ananassae. Biol Zentralblatt 1994; 113:469-474.

26. Karan D, Dahiya N, Munial AK et al. Desiccation and starvation tolerance of adult Drosophila: Opposite latitudinal clines in natural populations of three different species. Evolution 1998; 52:825-831.
27. Tishkoff SA, Varkonyi R, Cahinhinan N et al. Haplotype diversity and linkage disequilibrium at human G6PD: Recent origin of alleles that confer malarial resistance. Science Jul 20 2001; 293(5529):455-462.
28. Hamblin M, Di Rienzo A. Detection of the signature of natural selection in humans: Evidence from the Duffy blood group locus. Am J Hum Genet 2000; 66:1669-1679.
29. Hamblin MT, Thompson EE, Di Rienzo A. Complex signatures of natural selection at the Duffy blood group locus. Am J Hum Genet 2002; 70:369-383.
30. Rana BK, Hewett-Emmett D, Jin L et al. High polymorphism at the human melanocortin 1 receptor locus. Genetics Apr 1999; 151(4):1547-1557.
31. Hollox E, Poulter M, Zvarik M et al. Lactase haplotype diversity in the old world. Am J Hum Genet 2001; 68:160-172.
32. Mitchell-Olds. Arabidopsis thaliana and its wild relatives: A model system for ecology and evolution. Trends Ecol Evol 2001; 16:693-697.
33. Alonso-Blanco C, Koorneef M. Naturally occurring variation in Arabidopsis: An underexploited resource for plant genetics. Trends Plant Sci 2000; 5:22-29.
34. Abbott RJ, Gomes MF. Population genetic structure and outcrossing rate of Arabidopsis thaliana. Heredity 1989; 42:411-418.
35. Charlesworth D, Charlesworth B, Morgan M. The pattern of neutral molecular variation under the background selection model. Genetics 1995; 141:1619-1632.
36. Nordborg M, Charlesworth B, Charlesworth D. The effect of recombination on background selection. Genet Res Apr 1996; 67(2):159-174.
37. Bergelson J, Stahl E, Dudeck S et al. Genetic variation between and within populations of Arabidopsis thaliana. Genetics 1998; 148:1311-1323.
38. Miyashita NT, Kawabe A, Innan H. DNA variation in the wild plant Arabidopsis thaliana revealed by amplified fragment length polymorphism analysis. Genetics Aug 1999; 152(4):1723-1731.
39. Sharbel TF, Haubold B, Mitchell-Olds T. Genetic isolation by distance in Arabidopsis thaliana: Biogeography and postglacial colonization of Europe. Mol Ecol Dec 2000; 9(12):2109-2118.
40. Innan H, Stephan W. The coalescent in an exponentially growing metapopulation and its application to Arabidopsis thaliana. Genetics Aug 2000; 155(4):2015-2019.
41. Johanson U, West J, Lister C et al. Molecular analysis of fri, a major determinant of natural variation in Arabidopsis flowering time. Science 2000; 290:344-347.
42. Le Corre V, Roux F, Reboud X. DNA polymorphism at the FRIGIDA gene in Arabidopsis thaliana: Extensive nonsynonymous variation is consistent with local selection for flowering time. Mol Biol Evol 2002.
43. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics Nov 1989; 123(3):585-595.
44. McDonald JH, Kreitman M. Adaptive protein evolution at the Adh locus in Drosophila. Nature Jun 20 1991; 351(6328):652-654.
45. Pigliucci M. Ecological and evolutionary genetics of Arabidopsis. Trends Plant Sci Feb 1998; 3:485-489.
46. Barton NH. Linkage and the limits to natural selection. Genetics 1995; 140:821-841.
47. Clauss M, Cobban H, Mitchell-Olds T. Cross-species microsatellite markers for elucidating population genetic structure in Arabidopsis and Arabis (Brassicaceae). Mol Ecol 2002; 11:591-601.
48. Stranger B. Molecular Population Genetics of Arabidopsis species [PhD thesis]. Missoula: University of Montana, 2002.
49. Workman PL, Niswander JD. Population studies on southwestern Indian tribes. II. Local genetic differentiation in the Papago. Am J Hum Genet 1970; 22:24-49.
50. In: Balding JD, Bishop M, Cannings C, eds. Handbook of Statistical Genetics. Chichester, UK: Wiley, 2000.
51. Wright S. Evolution in Mendelian populations. Genetics 1931; 16:97-159.
52. Hudson R. A new statistic for detecting genetic differentiation. Genetics 2000; 155:2011-2014.

53. Kim Y, Stephan W. Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 2002; 160:765-777.
54. Fu YX. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics 1997; 147:915-925.
55. Crow JF. Basic Concepts in Population, Quantitative, and Evolutionary Genetics. New York: Freeman, 1986.
56. Maynard Smith J, Haigh J. The hitch-hiking effect of a favorable gene. Genet Res 1974; 23:23-35.
57. Ohta T, Kimura M. The effect of a selected linked locus on heterozygosity of neutral alleles (the hitchhiking effect). Genet Res 1975; 25:313-326.
58. Stephan W, Wiehe T, Lenz M. The effect of strongly selected substitutions on neutral polymorphism - analytical results based on diffusion theory. Theor Pop Biol 1992; 41:237-254.
59. Kaplan N, Hudson R, Langley C. The "hitchiking effect" revisited. Genetics 1989; 123:887-899.
60. An ancestral recombination graph. In: Griffiths RC, Marjoram P, Tavaré S, eds. Progress in population genetics and Human evolution. Berlin: Springer, 1997.
61. Barton NH. Genetic Hitch-hiking. Phil Trans R Soc Lond B 2000; 355:1553-1562.

INDEX

A

Accessory gland protein 16, 17
Adaptation 55, 79, 89, 107-110, 113, 114
Adaptive evolution 1, 15, 26, 65, 113, 114
Adaptive substitution 65, 74
Advantageous allele 35, 44, 48, 65, 104, 107, 109
Antigenic determinants 94, 95, 99, 100
Axoneme 24

B

Background selection 2, 8, 19, 22, 26-28, 35, 36, 43, 69, 74, 75, 104, 106-108, 111
Bayesian analysis 23, 24, 28, 29
Bottleneck 40-43, 48, 50, 51, 56, 58, 59, 61, 62, 69, 99

C

Change in population size 56, 73
Chloroquine resistance 100
Clines 105, 107, 109
Clonal population structure 99
Coadaptation 13-19
Coalescence 44, 48, 66, 68, 70, 75, 80, 82, 87
Coalescent 35, 39, 40, 45, 47, 48, 50, 55-57, 109, 111
Codon usage bias 26, 27, 98

D

Demographic history 55, 63, 75
Demographic sweep 98
Differentiation 17, 19, 73, 99, 100, 104, 106-108, 110-113
Distal conserved element (DCE) 25
DNA repeat region 99
DNA sequence polymorphism 1, 3, 5
Drosophila melanogaster 34, 66
Duffy factor 107
Dynein intermediate chain 22, 24

E

Ecological diversity 78, 79, 84
Ecotype 78-91
Effective population size 23, 26, 29, 36, 37, 60, 65, 67, 90, 106, 108, 110, 111
Emigrant haplotype 112
Environment 15, 17, 55, 89, 90, 95

F

Female sperm receptor 16
Frequency spectrum 22, 27, 28, 35, 36, 39, 41-45, 47, 48, 51, 67-69, 72-75, 107, 111
FRIGIDA (FRI) locus 109

G

Gene diversity 55-57, 60
Gene duplication 2-4, 27

H

Haplotype 1, 2, 7-10, 23, 34-45, 47-50, 52, 67, 68, 70-73, 107-109, 112
Haplotype distribution 34
Haplotype diversity 7, 34-36, 38, 39, 41, 72
Haplotype group 1, 7-10
Haplotype number 7, 34, 37, 38, 42-44
Haplotype partition 34, 37, 38
Haplotype structure 8, 10, 39, 41, 42, 44, 48, 49
Haplotype test 2, 7, 34, 37, 42, 45, 47, 48, 50, 52, 72
Heterozygosity 2, 42, 65-67, 69, 73, 110, 111
Hitchhiking 2, 3, 8, 9, 18, 34-36, 39-45, 47, 48, 50-52, 65-75, 104, 107, 108, 111, 114
HKA test 3, 66
Horizontal transfer 84

I

Immune selection 95, 99
Incomplete sweep 34, 36, 40
Infinite site model (ISM) 36, 37
Isolation-by-distance 106, 109

J

janus 3, 5

K

Ka/Ks 2, 5

L

Linkage 23, 26, 28, 36, 39, 44, 45, 51, 67, 70-73, 100, 104, 106
Linkage disequilibrium 36, 39, 44, 45, 70-73, 100, 106
lnRH 55-63
lnRV 55-63
Local adaptation 107-110, 113, 114
Local selective sweep 107-109, 113
Low frequency variant 3, 7, 35, 67, 74, 75

M

Major histocompatibility complex (MHC) 15
Malaria's eve 95-100
Male-female coadaptation 14, 15
Male-specific genes 24, 29
Metapopulation 109, 113
Microsatellite 55-60, 62, 63, 97, 100, 110
Migration 9, 43, 72, 73, 98, 104-106, 110, 112-114
Mitochondrial DNA (mtDNA) 96
MK test 3, 7
Molecular evolution 1, 3, 5, 13, 15, 17, 18, 29, 99, 106
Multilocus sequence typing (MLST) 87, 88

N

Neutrality 2, 3, 7, 8, 29, 34, 36, 38, 41, 44, 46, 47, 49-51, 56, 65, 67, 109, 114
Neutrality test 3, 7, 34, 36, 38, 41, 44, 46, 49, 51
Novel gene 22, 24, 30, 80
Number of polymorphic site 34, 37, 39, 40, 42, 44, 45, 107

O

ocnus 1, 3, 5, 23

P

P. falciparum 94-100, 107
Parasite-host 18
Periodic selection 78-91
Polymorphism 1-3, 5-10, 13-19, 22-24, 26-29, 34-36, 40, 41, 43, 45, 51, 56, 65, 67-69, 74, 75, 95-100, 106-109, 111, 112
Population expansion 3, 8, 41, 56, 58, 59, 62, 109, 113
Population genetics 108, 109
Population mutational parameter 34, 37, 40
Population subdivision 9, 23, 66, 72, 73, 110, 112
Positive selection 1-3, 5, 7, 8, 10, 15-18, 22, 23, 28, 29, 65, 71, 75, 96
Proximal conserved element (PCE) 25
Purifying selection 2, 7, 19, 66, 74

R

Rapid evolution 13-17, 19, 22, 29
Recombination 2, 3, 7, 8, 10, 18, 19, 22, 26, 34-36, 38-45, 47-51, 65-72, 74, 75, 78-84, 87-91, 98, 104-108, 110-114
Recombination in bacteria 78-80
Region of low recombination 26, 74, 106

S

Sdic 22, 24-30
Selective sweep 1-3, 8, 9, 18, 22-24, 26-30, 34-37, 40, 41, 43-48, 50, 51, 56, 59, 62, 63, 86-91, 99, 100, 106-109, 111-114
Sequence cluster 78, 82, 86, 90, 91
Sex-related gene 13, 15, 18, 19
Sexual conflict 13-16, 18, 19, 29
Sexual selection 2, 13, 14, 19, 29
Simulation 38-40, 42, 46-48, 50-52, 55-62, 66-68, 70, 71, 74, 75, 81, 82, 87, 88, 109-111
Speciation 2, 14, 19, 78, 79, 82-84, 99
Spermatogenesis 29, 30
Star clade 86, 87
Stepping stone model 112, 114
Structured population 49, 72, 104, 105, 107, 111, 114
Subpopulation 72, 73, 75, 104, 105, 110-114
Surface protein 15, 16, 95, 98, 99

T

Testis-specific element (TSE) 22, 25
Traffic 10, 36, 109, 110
Traffic model 10, 110
Type I error 55, 62, 63

W

Wright-Fisher model 36, 65, 67, 69

Z

Z_{nS} 38, 39, 45-48
Zona pellucida 17, 18